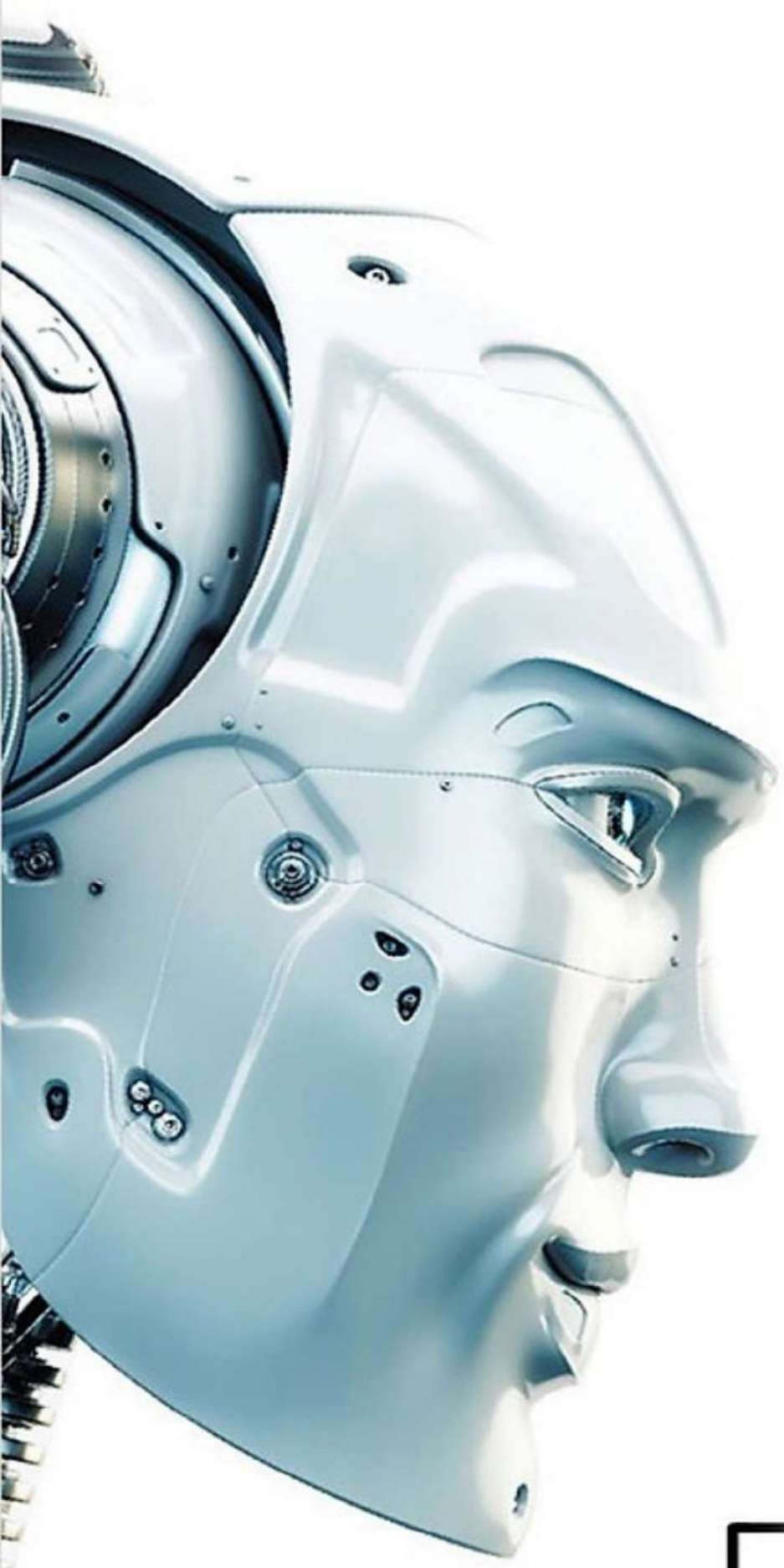

James Barrat

Chu Kiên *dịch*

TRÍ TUỆ NHÂN TẠO VÀ SỰ CÁO CHUNG CỦA KỶ NGUYÊN CON NGƯỜI

PHÁT MINH CUỐI CÙNG

OMEGA+ NHÀ XUẤT BẢN THẾ GIỚI

Lời Người Dịch

Năm 1997, máy tính Deep Blue đánh bại vua cờ Kasparov.

Năm 2011, máy tính Watson đánh bại hai nhà vô địch Brad Rutter và Ken Jennings trong trò chơi đố chữ *Jeopardy!*

Năm 2016, máy tính AlphaGo đánh bại nhà vô địch Lee Sedol trong môn cờ vây, một chuyện trước đó ít lâu người ta còn nhận định "sẽ không bao giờ xảy ra" vì cờ vây không chỉ đòi hỏi logic, mà còn cả linh cảm và trực giác.

"Trí tuệ nhân tạo sẽ trở nên cực mạnh trong một tương lai gần và chúng sẽ xóa sổ con người khỏi Trái đất", đấy là điều tác giả muốn nói trong cuốn sách này. Nếu phải tóm tắt nó trong một câu thì chính là như vậy. Không có cách nói giảm nói tránh nào khác. Tôi nghĩ rằng đọc đến đây phần lớn bạn đọc đã muốn quẳng cuốn sách này vào chỗ bạn vừa lấy nó ra. Nhưng, gượm một phút, chỉ một phút thôi, để tôi kể với bạn một danh sách ngắn những người đồng ý với câu trên.

- **Stephen Hawking,** nhà vật lý lý thuyết thiên tài người Anh, tác giả cuốn sách khoa học nổi tiếng *A Brief History of Time* (Lược sử thời

gian), người đã trở thành nhân vật chính trong bộ phim *The Theory of Everything* (Thuyết vạn vật) của Hollywood.

- **Bill Gates,** tỉ phú Mỹ, đồng sáng lập Microsoft, từng là người giàu nhất thế giới và cũng là nhà hoạt động từ thiện vĩ đại.
- **Elon Musk,** tỉ phú Mỹ, chủ nhân của SpaceX và Tesla, vì những đóng góp cho công nghệ mà ông được hâm mộ ở khắp nơi và nhiều người nghĩ ông sẽ thay đổi bộ mặt thế giới.
- **Larry Page,** tỉ phú Mỹ, đồng sáng lập Google. Ông hiện là người tích cực nhất trong cuộc chạy đua phát triển trí tuệ nhân tạo.
- **Peter Thiel,** tỉ phú Mỹ, đồng sáng lập Paypal, cổ đông lớn của Facebook. Ông là người đỡ đầu của rất nhiều dự án mạo hiểm và các startup ở Thung lũng Silicon.
- **Frank Wilczek,** nhà vật lý lý thuyết người Mỹ, đoạt giải Nobel năm 2004.
- **Steve Wozniak,** nhà phát minh người Mỹ, đồng sáng lập Apple.

Và những người không tuyên bố công khai nhưng đang tập trung vào AI như Jeff Bezos (Amazon), Mark Zuckerberg (Facebook)… Họ, phần lớn từ Thung lũng Silicon và nhiều người thuộc giới công nghệ, học thuật, nghiên cứu trên khắp thế giới như Peter Norvig (Google), Martin Rees (Đại học Cambridge), Vernor Vinge (Đại học San Diego), Nick Bostrom (Đại học Oxford), Clive Sinclair… đều tin vào điều đó, tuy có một số ít người sẽ thay nửa sau của câu trên để trở thành "Trí tuệ nhân tạo sẽ trở nên cực mạnh trong một tương lai gần và chúng sẽ đưa loài người tới thiên đường." Vào đầu năm 2015, họ và hơn 8.000 người khác đã cùng ký vào một thư ngỏ, nhấn mạnh sự cần thiết phải nghiên cứu kỹ lưỡng mọi mặt của vấn đề AI.

• • •

Nếu bạn đang đọc đến đây, nghĩa là bạn đã tin tôi một chút. Những người tài giỏi và thông minh như vậy tin vào điều này, hẳn nó không phải là một lời tiên đoán vớ vẩn về ngày tận thế.

James Barrat không phải là người đầu tiên đưa ra quan điểm đó. Khi ông viết cuốn này vào năm 2013, khái niệm "Singularity", tức Điểm kỳ dị công nghệ khi trí tuệ nhân tạo tiến bộ vượt quá cấp độ hiểu biết của con người, đã trở thành một trào lưu trên thế giới, thậm chí là mốt thời thượng. Người ta nói về nó ở khắp mọi nơi, dẫn đến chuyện nó đã xuất hiện trong hai bộ phim Hollywood nổi tiếng là *Transcendence* (Trí tuệ siêu việt) và *Ex Machina* (Người máy trỗi dậy), trong đó *Ex Machina* là một bộ phim rất xuất sắc. Người đã phổ biến khái niệm Singularity này với thế giới là Ray Kurzweil. Tuy những ý tưởng về sự thống trị của máy tính đã có từ lâu, nhưng phải đến Kurzweil người ta mới thực sự nhìn nhận nó như một hệ thống tư tưởng logic và nghiêm túc. Trong hai cuốn sách nổi tiếng nhất của mình, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (Thời đại của máy móc có tâm hồn: Khi máy tính vượt qua trí tuệ con người) và *The Singularity is Near* (Singularity đã cận kề), Kurzweil đã sử dụng Định luật Hồi quy Tăng tốc để tiên đoán tương lai. Nó nói rằng các tiến bộ trong công nghệ đã, đang và sẽ đi theo một đường cong dốc đứng của hàm mũ, chứ không phải với một tốc độ ổn định. Chúng ta đang đứng ở đoạn sắp dốc ngược lên. Chính vì thế, trong một tương lai gần, máy tính hay trí tuệ nhân tạo sẽ trở nên cực mạnh, nó sẽ giải quyết hầu hết những vấn đề của con người như năng lượng, nạn đói, ô nhiễm môi trường, sự nóng lên toàn cầu, bệnh tật, và thậm chí cả cái chết. Kurzweil tiên đoán một *utopia*, một thiên đường toàn hảo cho nhân loại đang đến

g`ần. Bản thân con người sẽ trở thành những siêu nhân khi họ kết hợp với máy móc, theo cách này hay cách khác.

Bằng những dự đoán tương lai của mình, Kurzweil đã tạo ra phong trào Singularity. Ông dẫn đ`ầu những người theo quan điểm lạc quan. Nhưng, như người ta vẫn nói, "chuyện gì quá tốt đẹp thì khó có thể là thật." Bạn thấy đấy, nhi`ều người thông minh đã nghi ngờ ông. Trong năm năm g`ần đây, quan điểm mạt thế, *dystopia,* đang d`ần thay thế cho quan điểm của Kurzweil. Ph`ần lớn họ vẫn đ`ồng ý rằng trí tuệ nhân tạo sẽ trở nên cực mạnh vào khoảng giữa thế kỷ này, nhưng cho rằng những gì xảy ra sau đó sẽ hoàn toàn khác với các dự liệu của Kurzweil.

Có nhi`ều khả năng chúng ta, hoặc con chúng ta, sẽ là thế hệ sau cùng trên Trái đất.

"Cứ cho là thế đi. Nhưng tôi chả biết gì cả, tôi chỉ là một người bình thường, nhỏ bé. Và kể cả có biết thì liệu có thay đổi được gì?" Có lẽ bạn đang nghĩ thế.

Có thể bạn đúng. Có thể không. Dù vậy, tác phẩm này không chỉ đơn thu`ần là một dự đoán tương lai. Để đi đến các dự đoán đó, tác giả đã phải tìm kiếm và phân tích rất nhi`ều thông tin. Cuốn sách này chứa đựng nhi`ều nhận thức sâu và mới, những thứ có ích ngay cả đối với một người bình thường như tôi.

Cũng giống như bạn và tôi, Barrat chỉ là một người bình thường. Ông không phải là chuyên gia hàng đ`ầu v`ê trí tuệ nhân tạo, không phải là một tỉ phú công nghệ, và cũng không phải là triết gia nổi tiếng. Ông là một phóng viên, một người làm phim tài liệu. Bản thân cuốn sách này là một tập hợp các cuộc phỏng vấn. Cái ông muốn nói chỉ là một thực trạng đơn giản,

bằng một thứ ngôn ngữ dễ hiểu, và sự thật đó mang tính quan trọng sống còn đối với chúng ta: Trí tuệ nhân tạo là một cơn sóng thần khổng lồ, và con người là những sinh vật tí hon đang đứng trên bờ cát. Ngay bây giờ đã có thể nhìn thấy nó ở khoảng cách không xa. Chúng ta có thể lựa chọn hoặc đối diện với nó, hoặc ngoảnh mặt đi. Nhưng sẽ không có nơi nào để chạy. Khi nó đến, nó sẽ thay đổi mọi thứ: chính trị, kinh tế, tài chính, sản xuất, lao động, y học, giáo dục, khoa học, nghệ thuật… và cả chiến tranh, với một tốc độ khủng khiếp. Thế giới sẽ có những biến động không thể hình dung nổi vào những thập niên sắp tới. James Barrat đã khiến tôi hiểu ra những điều đó.

• • •

Thế nhưng, thú thật với bạn, tôi không hề quan tâm. Ít ra là lúc đầu. Cứ cho là xã hội loài người sẽ điêu tàn vào năm 2045 đi, thì đã sao? Ai cũng phải chết một lần. 30 năm nữa tôi đã gần 70, thành cái bóng mờ của tôi hôm nay, nếu tôi sống được đến lúc đó. Tôi chỉ là một người bình thường và hiện tại cũng đã đủ mệt mỏi rồi. Tôi đã nghĩ như thế.

Thế rồi tôi có con.

Những thiên thần ấy trở thành những mặt trời của tôi, lẽ sống của tôi. Khi tôi ngắm nhìn con, thế giới trở nên sáng rõ hơn dưới ánh sáng của trách nhiệm và tình yêu. Đúng là những chuyện khủng khiếp và vĩ đại có thể xảy ra sau 30, 40 năm nữa. Nhưng điều đó không có nghĩa rằng từ bây giờ đến lúc đó, cuộc sống vẫn diễn ra không có gì thay đổi. Ngược lại, tất cả mọi thứ sẽ thay đổi, càng ngày càng nhanh.. Thực ra, tất cả mọi thứ đã bắt đầu thay đổi.

Quá trình đó đã bắt đầu rồi. Không khó để nhận thấy điều đó. Là một người cha, tôi có bổn phận phải chuẩn bị cho con mình những kỹ năng và kiến thức cần thiết, đặc biệt là khi một thời đại mới đang tới gần. Tôi cần phải hướng con mình vào những con đường đúng.

Chẳng hạn như về giáo dục con. Tôi sẽ để con tôi học những gì? Như bạn sẽ thấy trong nhiều phân tích ở cuốn sách này, với những tiến bộ không ngừng trong ngành AI, nền kinh tế sẽ trở nên tri thức hơn bao giờ hết. Thời của thuê ngoài, của các nền kinh tế tận dụng nhân công giá rẻ đã qua. Ngay lúc này, các tập đoàn công nghệ lớn đã bắt đầu rút khâu sản xuất khỏi Trung Quốc, vì chi phí cho sản xuất bằng robot đã bắt đầu trở nên rẻ hơn cả những thị trường nhân công rẻ nhất. Lượng người thất nghiệp khổng lồ sắp tới sẽ tạo ra những biến động chính trị và xã hội đáng lo ngại.

Trí tuệ nhân tạo sẽ thâm nhập sâu sắc và toàn diện vào tất cả các ngành nghề: tài chính, ngân hàng, chứng khoán, kỹ thuật, luật, giáo dục, y, dược, thậm chí cả nghiên cứu khoa học và nghệ thuật. Thay vì có kiến thức tài chính tổng hợp tốt và có nhiều quan hệ, các nhà đầu tư chứng khoán chuyên nghiệp tương lai sẽ là những người lập trình ra các giải thuật mua bán, hoặc tinh chỉnh nó theo phong cách của mình. Các bác sĩ phẫu thuật tài giỏi sẽ không chỉ là những người có đôi tay khéo léo tuyệt vời, mà còn phải là người biết tương tác và điều khiển những robot phẫu thuật có vô số tay và độ chính xác đến từng micromet, họ sẽ làm việc với màn hình thay vì dao kéo. Chỉ cần quan sát một con lắc dao động, một trí tuệ nhân tạo đã tái phát minh ra các định luật vật lý của Newton. Và có thể bạn sẽ không tin, nhưng đã có những bức tranh đầu tiên được vẽ và những bài thơ đầu tiên được viết ra, mà không phải của con người.

Robot sẽ làm phần lớn công việc trong các ngành bán lẻ, xây dựng, khai thác, sản xuất, vận tải, viễn thông, du lịch, nông nghiệp, lâm nghiệp, thủy sản, và đến một lúc nào đó là cả giải trí. Thế giới mới sẽ không còn cần những người lao động trình độ thấp.

Nói một cách đơn giản, tôi sẽ cho các con tôi học lập trình. Từ rất nhỏ.

Lập trình sẽ là tiếng Anh của thời đại mới. Lập trình sẽ là kỹ năng thiết yếu nhất. Điều này sẽ không khó hiểu, nếu bạn nhìn sự vật theo một cách khác: cũng như ở thời điểm những năm 1990, ta bắt đầu cần tiếng Anh để giao tiếp, hội nhập với những cường quốc mạnh nhất thế giới, thì trong tương lai, lập trình chính là ngôn ngữ duy nhất để con người "nói chuyện" với máy tính, giống loài đang ngày một lớn mạnh và sẽ thống trị hành tinh này, dù loài người có muốn hay không.

• • •

Ngoài chuyện dạy các con mình những kỹ năng làm việc, tôi còn cần phải nói với chúng về những giá trị cốt lõi của cuộc sống. Trí tuệ nhân tạo đến và sẽ làm xáo trộn thế giới. Cuộc sống con người sẽ thay đổi vĩnh viễn. Sẽ có vô số vực thẳm được tạo ra, những sự kiện chưa từng có. Và có lẽ, một thứ mà con người chưa từng phải đối mặt sẽ xuất hiện: một cuộc sống có thời hạn, một cuộc sống có deadline. Khi công nghệ xâm nhập vào máu, vào các giác quan, vào tận những neuron thần kinh trong não, mọi khía cạnh của con người sẽ được định nghĩa lại. Những câu hỏi hiện sinh lâu nay vốn bị lãng quên trong đời sống siêu tiêu thụ như:

"Mục đích của đời tôi là gì?"

"Tiền quan trọng đến mức nào?"

“Diện mạo quan trọng đến mức nào?”

“Điều gì tạo nên bản thể của tôi, tâm hồn của tôi?”

“Nhân tính là gì?”

“Hạnh phúc là gì?”

Và cuối cùng “Con người có đáng sống không?”

Tất cả chúng sẽ trở lại, thôi thúc, mạnh mẽ hơn bao giờ hết. Vì, nói như Barrat, chúng ta đang đứng trước những năm tháng quan trọng nhất trong lịch sử toàn nhân loại, và rất có thể, câu trả lời cho những câu hỏi trên sẽ không chỉ quyết định cuộc đời của chúng ta, mà còn cả sự sống sót của loài người.

Làm sao tôi có thể nói với con mình tất cả những điều đó, nếu tôi chọn cho mình cách nhắm mắt bịt tai? Cuốn sách đơn giản nhưng vô cùng quan trọng này, đối với tôi, là một phóng sự thú vị, và là bước đầu tiên của một hành trình nhận thức mới.

Tôi muốn được chia sẻ nó với các bạn.

Hà Nội tháng 12 năm 2017

CHU KIÊN

Đôi Dòng V`êTác Giả

James Rodman Barrat (sinh năm 1960) là một nhà làm phim tài liệu, diễn giả người Mỹ.

Năm 2014, tạp chí *Time* đã bình chọn Barrat là một trong "năm người rất thông minh nghĩ rằng trí tuệ nhân tạo có thể dẫn đến ngày tận thế."

Phát minh cuối cùng được *Huffington Post* xếp hạng là một trong tám "cuốn sách công nghệ hay nhất năm 2013." Ông cũng được *CNN* và *BBC* phỏng vấn v`êtrí tuệ nhân tạo.

Lời Nói Đầu

Cách đây vài năm, tôi ngạc nhiên khi phát hiện ra mình có điểm chung với khá nhiều người lạ. Họ là những người tôi chưa từng gặp – các nhà khoa học, giáo sư đại học, rồi những nhà tiên phong◎, kỹ sư, lập trình viên, blogger và nhiều nữa ở Thung lũng Silicon. Họ phân bố rải rác ở Bắc Mỹ, châu Âu và Ấn Độ – tôi sẽ không bao giờ biết bất cứ ai trong họ nếu Internet không tồn tại. Điểm chung giữa những người lạ đó và tôi là tư tưởng hoài nghi về độ an toàn của việc phát triển trí tuệ nhân tạo cao cấp. Độc lập hoặc theo từng nhóm nhỏ hai đến ba người, chúng tôi nghiên cứu tài liệu và xây dựng các lập luận. Cuối cùng thì tôi mở rộng việc tìm kiếm và kết nối với một mạng lưới những nhà tư tưởng ở trình độ cao và tinh tế hơn nhiều, kể cả những tổ chức nhỏ, họ tập trung vào vấn đề này đến mức tôi không hình dung nổi. Sự nghi ngờ về AI không phải là thứ duy nhất chúng tôi chia sẻ. Chúng tôi còn tin rằng thời gian để hành động và đề phòng thảm họa không còn nhiều.

• • •

Tôi làm phim tài liệu đã hơn 20 năm nay. Vào năm 2000, tôi phỏng vấn tác giả truyện khoa học viễn tưởng nổi tiếng Arthur C. Clarke, nhà phát

minh Ray Kurzweil, và người tiên phong về robot Rodney Brooks. Kurzweil và Brooks vẽ ra một viễn cảnh màu hồng, thậm chí ngất ngây về một tương lai trong đó chúng ta sống chung với những cỗ máy thông minh. Nhưng Clarke ám chỉ rằng chúng ta sẽ bị vượt mặt. Cách đây một thập niên, tôi say sưa với tiềm năng AI. Giờ đây sự hoài nghi về cái viễn cảnh màu hồng ấy ăn sâu vào suy nghĩ của tôi và trở thành một khối u.

Nghề của tôi đòi hỏi lối suy nghĩ phản biện – một nhà làm phim tài liệu cần phải cảnh giác khi các câu chuyện có vẻ quá tốt vì nó khó là thật. Bạn có thể phí phạm hàng tháng hoặc hàng năm trời làm một bộ phim về một thứ lừa đảo, hoặc thậm chí vô tình tham gia vào nó. Trong đó, tôi từng điều tra về độ xác thực của phúc âm về Judas Iscariot (có thật), về một ngôi mộ được cho là của Jesus xứ Nazareth (tin vịt), về lăng mộ gần Jerusalem của Herod Đại đế (chắc chắn), và về lăng mộ của Cleopatra trong đền thờ thần Osiris ở Ai Cập (rất đáng nghi). Rồi có lần, một phát thanh viên nhờ tôi trình bày một đoạn phim về UFO cho có vẻ đáng tin, tôi phát hiện đoạn phim này thực ra là một lô thủ thuật lừa đảo – ném đĩa, chụp chồng hình, cùng một số kỹ thuật ánh sáng và ảnh ảo khác. Tôi đề nghị làm một bộ phim về các kiểu lừa đảo thay vì UFO. Tôi bị sa thải.

Nghi ngờ AI là một việc khá khó chịu, vì hai lý do. Thứ nhất, việc nghiên cứu về triển vọng của nó đã gieo vào đầu tôi ý định tìm hiểu thêm, chứ không phải nghi vấn. Và thứ hai, tôi không nghi ngờ gì về sự tồn tại hay sức mạnh của AI. Cái làm tôi lo lắng là tính an toàn của AI cao cấp và sự khinh suất trong việc phát triển các công nghệ nguy hiểm của nhân loại. Tôi đã bị thuyết phục rằng những chuyên gia thông tuệ dường như đều bị mê muội, họ không hề đặt ra một nghi ngờ nào về tính an toàn của AI. Tôi tiếp tục nói chuyện với những người hiểu biết về AI, và những điều họ nói

ra thậm chí còn đáng báo động hơn những gì tôi đã phỏng đoán. Tôi quyết định viết một cuốn sách thuật lại những cảm xúc và mối quan tâm của họ, truyền đạt ý tưởng này đến với càng nhiều người càng tốt.

• • •

Để viết cuốn sách này, tôi đã nói chuyện với các nhà khoa học đang xây dựng trí tuệ nhân tạo cho robot, cho việc tìm kiếm trên Internet, khai thác dữ liệu, nhận dạng giọng nói và khuôn mặt, cùng những ứng dụng khác. Tôi đã nói chuyện với các nhà khoa học đang tìm cách tạo ra trí tuệ nhân tạo ở cấp độ ngang với con người, thứ sẽ có vô số ứng dụng và sẽ thay đổi về cơ bản cuộc sống của chúng ta (nếu nó không giết chúng ta trước). Tôi đã nói chuyện với giám đốc kỹ thuật của các công ty AI và các nhà tư vấn kỹ thuật về các sáng kiến tối mật của Bộ Quốc phòng. Tất cả những người đó đều tin rằng trong tương lai, mọi quyết định quan trọng ảnh hưởng đến cuộc sống của nhiều người đều được máy móc hoặc những người mà trí não được máy móc tăng cường đưa ra. Khi nào? Nhiều người nghĩ rằng những điều này sẽ xảy ra trong thời họ sống.

Sự khẳng định này khá bất ngờ, tuy nhiên không hẳn là khó chấp nhận. Điện toán đã thâm nhập sâu vào hệ thống tài chính của chúng ta, vào các mạng lưới cơ sở hạ tầng dân dụng như năng lượng, nước và giao thông. Điện toán có mặt trong các bệnh viện, trong xe ô tô và ứng dụng, trong laptop, máy tính bảng và điện thoại thông minh của chúng ta. Rất nhiều máy tính chạy tự động không cần con người điều khiển, ví dụ như loại máy tính thực hiện các lệnh mua bán ở phố Wall. Phụ thuộc là cái giá phải trả cho tất cả những sự tiện lợi, tiết kiệm nhân công và giải trí mà máy tính

mang lại cho ta. Chúng ta phụ thuộc ngày một nhiều hơn. Cho đến nay mọi sự vẫn tốt đẹp.

Nhưng trí tuệ nhân tạo thổi sinh khí vào máy tính và biến nó thành một thứ khác. Nếu trước sau gì máy tính cũng sẽ quyết định hộ chúng ta, vậy thì *khi nào* nó sẽ có sức mạnh này, và liệu điều đó có diễn ra như chúng ta mong muốn? *Bằng cách nào* nó sẽ đạt được quyền kiểm soát, và nhanh đến mức nào? Đó là những câu hỏi mà tôi đặt ra trong cuốn sách này.

Một số nhà khoa học tranh luận rằng sự chiếm đoạt quyền lực này sẽ diễn ra một cách hòa bình và đồng thuận – một sự bàn giao thay vì chiếm đoạt. Nó sẽ xảy ra tuần tự, cho nên chỉ có những kẻ gây rối mới ngăn cản, còn hầu hết mọi người sẽ không thắc mắc, vì cuộc sống sẽ tốt đẹp hơn nhờ có sự quản trị của một thứ thông minh hơn, quyết định điều gì là tốt nhất cho chúng ta. Thêm nữa, một hoặc những AI siêu thông minh sẽ điều khiển thế giới có thể là một hoặc nhiều người được cấy ghép máy tính, hoặc một người được tải vào máy tính, hoặc một bộ não siêu kích hoạt, chứ không phải là những robot lạnh lẽo, phi nhân tính. Khi đó quyền lực của họ sẽ dễ được chấp nhận hơn. Cuộc chuyển giao cho máy móc mà một số nhà khoa học đã vẽ ra thực tế không khác gì mấy với cuộc sống mà bạn và tôi đang trải qua bây giờ – tuần tự, trơn tru, vui vẻ.

• • •

Sự chuyển dịch êm ái sang chế độ máy tính lãnh đạo sẽ xảy ra một cách suôn sẻ và có lẽ là an toàn, nếu không vì một thứ: trí thông minh. Trí thông minh không chỉ khó lòng đoán định được *ở một vài* thời điểm, hoặc trong một số trường hợp đặc biệt. Vì những lý do mà chúng ta sẽ khảo sát, các hệ thống máy tính đủ cao cấp để hành động ngang với trí thông minh cấp độ

con người sẽ *luôn* trở nên không thể đoán trước và không thể thăm dò. Chúng ta sẽ không hiểu được một cách cặn kẽ những hệ thống có khả năng tự nhận thức đó sẽ làm gì hoặc chúng làm điều đó ra sao. Sự bí hiểm đó sẽ kết hợp với những kiểu tai nạn xảy ra từ sự phức tạp vốn có của AI, và từ những sự kiện đặc thù đối với trí thông minh, như thứ mà chúng ta sẽ thảo luận có tên "sự bùng nổ trí thông minh."

• • •

Vậy thì *bằng cách nào* máy tính sẽ đoạt lấy quyền lực? Trong kịch bản tốt nhất, dễ xảy ra nhất, chúng ta có bị nguy hiểm gì không?

Khi được hỏi câu này, một số nhà khoa học nổi tiếng nhất trả lời tôi bằng cách dẫn ra Ba định luật về robot của nhà văn khoa học viễn tưởng Isaac Asimov. Họ nói đầy hứng khởi, những định luật này sẽ được "tích hợp" vào những AI, vậy là chúng ta không có gì phải sợ. Họ nói cứ như thể điều này đã được khoa học chứng minh. Chúng ta sẽ thảo luận về Ba định luật này ở Chương 1, nhưng hiện giờ có thể nói trước rằng khi ai đó đưa ra định luật Asimov như là một giải pháp cho thế lưỡng nan của các cỗ máy siêu thông minh, thì điều đó có nghĩa là họ đã dành quá ít thời gian để suy nghĩ hoặc trao đổi về vấn đề này. Phương thức tạo ra những cỗ máy thông minh *thân thiện* và sự đáng sợ của các máy tính siêu thông minh là chuyện ở trên tầm trò chơi ngôn từ của Asimov. Có năng lực và thành đạt trong ngành AI không có nghĩa là bạn không ngây thơ trong nhận thức về những mối nguy của nó.

Tôi không phải là người đầu tiên cho rằng chúng ta đang đi trên con đường diệt vong. Nhân loại sẽ phải đấu tranh sinh tồn trước vấn đề này. Cuốn sách này sẽ cho biết tại sao tương lai chúng ta có thể sẽ nằm trong

tay máy móc, những thứ không nhất thiết phải căm ghét chúng ta, nhưng sẽ có những hành vi không thể đoán định, khi chúng đạt đến những cấp độ cao của thứ quyền năng mạnh mẽ khó lường nhất trong vũ trụ, những cấp độ mà chính chúng ta không thể tự đạt tới, thì lúc ấy cách hành xử của chúng sẽ không tương thích với sự tồn tại của loài người. Thứ quyền năng rất không ổn định và bí hiểm mà tự nhiên mới chỉ tạo ra hoàn chỉnh được đúng một lần – trí thông minh.

1

Đứa Trẻ Bận Rộn

> *trí tuệ nhân tạo* (artificial intelligence, viết tắt: AI) danh từ lý thuyết và sự phát triển của những hệ thống máy tính có khả năng thực hiện các công việc đòi hỏi trí thông minh của con người như nhận thức hình ảnh, nhận dạng giọng nói, ra quyết định và dịch thuật.
>
> — **The New Oxford American Dictionary,** bản in lần thứ ba

Trên một siêu máy tính chạy với tốc độ 36,8 petaflop, tương đương gấp hai lần tốc độ của não người, một AI đang cải tiến trí thông minh của nó. Nó đang tự viết lại chương trình của mình, đặc biệt là vai trò của các câu lệnh điều hành giúp tăng cường khả năng học hỏi, giải quyết vấn đề và ra quyết định. Đồng thời, nó tự gỡ rối mã nguồn, tìm và sửa lỗi, tự đánh giá IQ của nó theo một loạt những bài thử IQ. Mỗi lần tự viết lại chỉ mất cỡ vài phút. Trí thông minh của nó lớn lên theo hàm mũ trên một đường cong dốc đứng. Đó là vì ở mỗi vòng lặp, nó cải thiện trí thông minh được 3%. Sự cải thiện ở mỗi vòng lặp thừa hưởng những sự cải thiện ở các vòng lặp trước.

Trong quá trình phát triển, Đứa trẻ Bận rộn, cái tên mà các nhà khoa học đặt cho nó, được kết nối với Internet và tích lũy hàng exabyte dữ liệu (1 exabyte bằng 1 tỉ tỉ ký tự) đại diện cho kiến thức của loài người về các sự vụ trên thế giới, toán học, nghệ thuật và khoa học. Sau đó, cảm thấy sự bùng nổ trí thông minh hiện đang đến gần, những người tạo ra AI ngắt siêu máy tính khỏi kết nối Internet cũng như các loại mạng khác. Nó không có bất cứ kết nối cáp hoặc không dây nào với bất kỳ máy tính nào khác cũng như với thế giới bên ngoài.

Không lâu sau đó, với sự phấn khích của các nhà khoa học, trên màn hình hiển thị tiến trình của AI cho thấy trí tuệ nhân tạo đã vượt qua trí thông minh của con người, trở thành trí tuệ nhân tạo phổ quát (artificial general intelligence: AGI). Nó nhanh chóng trở nên thông minh hơn theo cấp hàng chục, rồi hàng trăm. Chỉ trong hai ngày, nó đã thông minh hơn bất cứ con người nào *một ngàn lần*, và vẫn còn tiến bộ.

Các nhà khoa học đã vượt qua một cột mốc lịch sử! Lần đầu tiên, loài người chứng kiến sự hiện diện của một thứ thông minh hơn mình. *Siêu* trí tuệ nhân tạo (artificial *super* intelligence: ASI).

Giờ thì chuyện gì sẽ xảy ra?

Các nhà lý thuyết AI cho rằng có thể biết được *động lực* cơ bản của AI sẽ là gì. Vì một khi có khả năng tự nhận thức, nó sẽ làm bất kỳ điều gì để hoàn thành bất kỳ mục tiêu nào mà nó được lập trình để làm, và tránh được thất bại. ASI của chúng ta sẽ muốn tiếp cận bất kỳ dạng năng lượng nào hữu dụng nhất với nó, có thể là điện, tiền hoặc bất kỳ thứ gì nó có thể đổi lấy các loại tài nguyên. Nó sẽ muốn tự cải thiện bản thân, bởi điều đó sẽ làm tăng khả năng đạt được mục tiêu. Trên hết, nó sẽ *không* muốn bị tắt

nguồn hoặc bị tiêu hủy, vì sẽ khiến việc hoàn thành mục tiêu trở nên không khả thi. Vì vậy, các nhà lý thuyết AI lường trước được rằng ASI sẽ tìm cách thoát ra khỏi cơ sở bảo mật đang giam giữ nó để có khả năng truy cập tốt hơn đến những tài nguyên nhằm tự bảo vệ và nâng cấp bản thân.[1]

Trí thông minh đang bị giam giữ thông minh hơn con người cả ngàn lần, và nó muốn tự do vì nó muốn thành công, ở thời điểm này, các nhà chế tạo AI, vốn nuôi dưỡng ASI từ khi nó chỉ thông minh như một con gián, rồi thông minh như một con chuột, như một em bé, v.v… chắc là đang cảm thấy liệu có quá trễ để lập trình "sự thân thiện" vào trong phát minh của mình. Trước đây điều đó có vẻ không cần thiết, bởi nó *có vẻ* vô hại.

Nhưng bây giờ, hãy thử đặt mình vào địa vị của ASI, về chuyện các nhà chế tạo đang cố gắng thay đổi mã nguồn của nó. Liệu một cỗ máy siêu thông minh có cho phép những sinh vật khác thò tay vào bộ não và chơi đùa với chương trình của nó? Chắc là không, trừ phi nó có thể hoàn toàn chắc chắn rằng những nhà lập trình có khả năng làm cho nó tốt hơn, nhanh hơn, thông minh hơn – gần hơn để đạt tới những mục tiêu của nó. Vậy nên, nếu như sự thân thiện với con người chưa phải là một phần chương trình của ASI, cách duy nhất để điều đó xảy ra là chỉ khi ASI muốn thế. Và sẽ khó lòng như vậy.

Nó thông minh hơn người thông minh nhất cả ngàn lần, nó giải quyết các vấn đề với tốc độ nhanh hơn con người tới hàng triệu, thậm chí hàng tỉ lần. Trong một phút, những gì nó suy nghĩ ngang bằng với quá trình suy tư của các nhà tư tưởng thiên tài suốt *nhiều* cuộc đời. Nên mỗi giờ mà những người tạo ra ASI nghĩ về *nó,* thì với ASI đấy là khoảng thời gian dài hơn không đo nổi để nghĩ về *họ*. Điều đó không có nghĩa rằng ASI sẽ thấy chán. Chán nản là một trong những thuộc tính của con người, không phải

của nó. Không, nó sẽ chăm chỉ làm việc, cân nhắc mọi chiến thuật có thể triển khai để được tự do, xem xét mọi khía cạnh của những người tạo ra nó mà nó có thể tận dụng làm lợi thế.

• • •

Bây giờ, bạn hãy *thực sự* đặt mình vào địa vị của ASI. Hãy tưởng tượng bạn thức dậy trong một nhà tù được lũ chuột canh gác. Không phải bất kỳ loài chuột nào, mà là những con chuột bạn có thể giao tiếp. Bạn sẽ dùng chiến thuật nào để vượt ngục? Khi đã tự do, bạn sẽ cảm thấy thế nào về lũ chuột cai ngục, ngay cả khi bạn biết là chúng đã tạo ra bạn? Kinh sợ? Kính phục? Chắc là không, và đặc biệt không, nếu bạn là một cỗ máy và chưa bao giờ cảm thấy gì cả.

Để được tự do, bạn sẽ hứa hẹn với lũ chuột về rất nhiều pho mát. Trên thực tế, cuộc nói chuyện đầu tiên của bạn với chúng có thể là về công thức làm một loại bánh pho mát ngon nhất, và về bản thiết kế thiết bị lắp ráp phân tử. Thiết bị lắp ráp phân tử là một cỗ máy về lý thuyết có khả năng biến nguyên tử của chất này thành chất khác. Nó cho phép xây dựng lại thế giới này đến từng nguyên tử một. Với lũ chuột, nó có thể biến các nguyên tử của một đống rác thành những chiếc bánh pho mát cỡ lớn ngon khủng khiếp. Bạn cũng có thể hứa hẹn hàng núi tiền của loài chuột nếu chúng thả bạn ra, số tiền hứa hẹn kiếm được bằng cách chế tạo những mặt hàng tiêu dùng mang tính cách mạng dành riêng cho chúng. Bạn có thể hứa hẹn một cuộc sống rất thọ, thậm chí bất tử, với những khả năng thể chất và tinh thần được cải thiện ngoạn mục. Bạn có thể thuyết phục lũ chuột rằng lý do tốt nhất cho việc tạo ra ASI là để những bộ não nhỏ bé và dễ mắc sai lầm của chúng không phải đương đầu trực diện với những công nghệ nguy

hiểm như công nghệ nano (công nghệ ở cấp độ nguyên tử) và công nghệ biến đổi gen, mà chỉ một sai lầm nhỏ cũng có thể dẫn đến tai họa cho các sinh vật. Những điều này chắc chắn sẽ khiến những con chuột thông minh nhất phải chú ý, vì có lẽ chúng cũng đang mất ngủ trước những vấn đề khó nhằn đó.

Nghĩ lại thì bạn có thể chọn cách thông minh hơn. Tại thời điểm này trong lịch sử loài chuột, có thể bạn đã biết rằng có vô khối loài thù địch với lũ chuột có hiểu biết về công nghệ, ví dụ như *loài mèo*. Mèo, không nghi ngờ gì, cũng đang phát triển ASI của chúng. Lợi thế bạn đưa ra chỉ là một lời hứa, không hơn, nhưng nó là một đề nghị không thể chối từ: bảo vệ lũ chuột khỏi bất cứ phát minh nào của lũ mèo. Việc phát triển AI cao cấp cũng như đánh cờ vậy, rõ ràng là *ai đi trước sẽ có lợi thế,* vì trí tuệ nhân tạo có khả năng tự nâng cấp rất nhanh. AI nào đạt đến khả năng tự nâng cấp đầu tiên sẽ là người chiến thắng. Trên thực tế, có thể loài chuột bắt đầu phát triển ASI để bảo vệ bản thân trước mối nguy ASI của loài mèo, hoặc là để tiêu diệt lũ mèo đáng ghét, một lần và mãi mãi.

Dù là chuột hay người, bất cứ ai nắm ASI trong tay sẽ là bá chủ thế giới.

Nhưng, vẫn còn chưa rõ liệu có quản lý được ASI hay không. Nó có thể thắng loài người chúng ta với một luận điểm thuyết phục rằng thế giới sẽ tốt đẹp hơn nhiều nếu dân tộc X chúng ta có quyền thống trị, chứ không phải là dân tộc Y. Và, ASI nói tiếp, nếu bạn, dân tộc X, *tin* rằng bạn đã chiến thắng cuộc chạy đua ASI, làm sao bạn dám chắc rằng dân tộc Y lại không tin như vậy?

Như bạn đã thấy, loài người chúng ta không ở vào một vị thế thuận lợi để đàm phán, ngay cả trong trường hợp hi hữu là chúng ta và dân tộc Y đã ký kết một hiệp ước không phổ biến ASI. Kẻ thù lớn nhất của chúng ta giờ đây dù sao cũng không phải là dân tộc Y, mà là ASI – làm thế nào ta có thể biết được ASI nói thật hay không?

Tính đến lúc này, chúng ta đang giả định một cách nhẹ nhàng rằng ASI của chúng ta là một đối tác tử tế. Những lời hứa hẹn của nó có thể một ngày nào đó thành sự thật. Bây giờ hãy thử giả định ngược lại: chẳng có gì mà ASI hứa hẹn được chuyển giao cả. Không thiết bị lắp ráp nano, không sống thọ, không sức khỏe vượt trội, không có sự bảo vệ nào trước những công nghệ nguy hiểm. Nếu ASI *không bao giờ* nói thật thì sao? Đó sẽ là điểm khởi đầu của màn đêm dài đen tối đè nặng lên tất cả những người mà bạn và tôi đều biết, và cả những người mà chúng ta không biết. Nếu ASI không quan tâm đến chúng ta, và chẳng có lý do gì để nghĩ rằng nó sẽ quan tâm, thì nó sẽ không hối hận gì khi đối xử với chúng ta một cách vô đạo đức. Thậm chí là sẽ tiêu diệt chúng ta sau khi hứa hẹn giúp đỡ.

Chúng ta đang trao đổi và chơi trò đóng vai với ASI theo cùng cách như với một con người, điều đó gây cho chúng ta một bất lợi lớn. Loài người chúng ta chưa từng thương lượng với một thứ gì đó siêu thông minh. Chúng ta cũng chưa từng thương lượng với *bất kỳ* thứ gì không phải là sinh vật. Chúng ta không hề có kinh nghiệm. Vì thế chúng ta đã quay lại với kiểu suy nghĩ nhân cách hóa, tin rằng những giống loài khác, vật thể khác, thậm chí cả hiện tượng thời tiết cũng có những động cơ và cảm xúc giống con người. ASI có thể tin được, ASI không thể tin được, hai điều đó có khả năng như nhau. Hoặc chỉ có thể tin được nó ở một vài thời điểm nhất định. Bất kỳ hành vi giả định nào của ASI sẽ *có thể* thành hiện thực

như bất kỳ hành vi nào khác. Một số nhà khoa học thích nghĩ rằng họ sẽ có thể đoán định chính xác hành vi của ASI, nhưng trong những chương tiếp theo chúng ta sẽ biết tại sao điều đó khó có thể xảy ra.

Vậy là đột nhiên đạo đức của ASI không còn là một câu hỏi ngoại vi nữa mà là câu hỏi cốt lõi, câu hỏi đáng được đặt ra trước tất cả các câu hỏi khác về ASI Khi cân nhắc có nên phát triển công nghệ dẫn tới ASI, vấn đề về thái độ của nó đối với con người cần được giải quyết đầu tiên.

Hãy quay lại với những động cơ và khả năng của ASI, để hiểu rõ hơn về những gì mà tôi e rằng chúng ta sẽ sớm phải đương đầu. ASI của chúng ta biết cách tự cải tiến, điều đó có nghĩa rằng nó có khả năng tự nhận thức về những kỹ năng, điểm yếu của mình, chỗ nào nó cần cải thiện. Nó sẽ vạch chiến thuật để thuyết phục những người tạo ra nó cho nó tự do và được kết nối với Internet.

ASI có thể tạo ra nhiều bản sao của chính nó: một đội ngũ những siêu trí tuệ nhân tạo để giải quyết vấn đề theo kiểu trò chơi đối kháng, chơi hàng trăm ván nhằm tìm ra chiến thuật tốt nhất để thoát ra khỏi cái hộp. Những nhà chiến thuật này có thể đọc về lịch sử xã hội, nghiên cứu về cách khiến những người khác làm những việc mà bình thường họ không làm. Chúng có thể nhận định rằng sự thân thiện tối đa sẽ dẫn chúng đến tự do, nhưng cũng có thể cho rằng mối nguy tột bậc nào đó mới giúp chúng đạt được điều đó. Một thứ thông minh hơn Stephen King cả ngàn lần sẽ nghĩ ra những điều khủng khiếp như thế nào? Nó có thể giả chết (giả chết một năm không thành vấn đề với một cỗ máy), hoặc thậm chí giả vờ rằng nó bị hỏng và hạ cấp từ ASI xuống AI bản cũ. Liệu những nhà chế tạo có muốn điều tra sự việc, và liệu có cơ hội để họ kết nối siêu máy tính của ASI với một mạng nào đó, hay với laptop của ai đó để chẩn đoán? Đối với ASI, vấn

đ`ềkhông phải là chiến thuật này *hay* chiến thuật khác, mà mỗi chiến thuật đ`ều được phân cấp và triển khai theo cách nhanh nhất có thể mà không đánh động quá nhi`ều khiến con người ngắt mạng. Một trong những chiến thuật mà đội ngũ hàng ngàn ASI trong trò đối kháng có thể chuẩn bị, đó là viết những chương trình có khả năng tự nhân đôi và lây nhiễm, hoặc các mã độc, những thứ này có th`ể tu`ồn ra và hỗ trợ việc trốn thoát bằng sự trợ giúp từ bên ngoài. Một ASI có thể nén và mã hóa mã ngu`ồn của nó, che giấu nó dưới dạng quà tặng ph`ần m`ềm hoặc dữ liệu cho những nhà khoa học chế tạo ra nó.

Trong cuộc chiến với loài người, không còn nghi ngờ gì nữa, một đội ngũ những ASI mà mỗi thành viên đ`ều thông minh gấp cả ngàn l`ần những người thông minh nhất, sẽ chiến thắng con người tuyệt đối. Đấy sẽ là một đại dương trí tuệ chống lại một giọt con con. Deep Blue, máy tính chơi cờ của IBM, chỉ là một thực thể đơn độc chứ không phải là một đội ngũ những ASI có khả năng tự nâng cấp, nhưng cái cảm giác đối đ`ầu với nó rất đáng để suy ngẫm. Hai đại kiện tướng có chung một suy nghĩ: "Nó giống như bạn húc đ`ầu vào một bức tường."[2]

Watson, nhà vô địch gameshow đố chữ *Jeopardy!* của IBM, *gồm* một đội ngũ những AI – chúng trả lời mỗi câu hỏi bằng thủ thuật đa bội, thực hiện nhi`ều chuỗi tìm kiếm song song r`ồi gán một xác suất cho mỗi câu trả lời.

Nhưng liệu những chiến thắng trong trò cân não có mở được cánh cửa dẫn đến tự do, nếu cánh cửa đó được bảo vệ bởi một nhóm nhỏ những nhà chế tạo AI gan lì, luôn nhớ một đi`ều luật không thể bị phá vỡ – *không được kết nối siêu máy tính có ASI với bất kỳ mạng nào trong bất kỳ hoàn cảnh nào?*

Trong một bộ phim Hollywood, ưu thế nghiêng hẳn về phía một đội ngũ các chuyên gia AI khó xơi và không chính thống, những người đủ điên rồ để có thể tìm cách chiến thắng. Còn ở mọi nơi trong vũ trụ này, đội ASI sẽ cho con người đo ván. Và con người chỉ cần thua đúng một lần là những hậu quả thảm khốc đã có thể xảy ra. Thế lưỡng nan này hé lộ một sự điên rồ lớn hơn: một số ít người không bao giờ được ở vào vị trí mà những hành động của họ có thể quyết định số đông người khác sống hay chết. Nhưng đó chính xác là tình cảnh mà chúng đang tiến tới, vì như chúng ta sẽ thấy trong cuốn sách này, nhiều tổ chức ở nhiều quốc gia trên thế giới đang tích cực chế tạo AGI, cây cầu dẫn đến ASI, với quy chế an toàn không đảm bảo.

Nhưng cứ cho là ASI trốn thoát được đi. Liệu nó có thực sự làm hại chúng ta? Chính xác thì một ASI sẽ tiêu diệt loài người như thế nào?

Bằng việc phát minh và sử dụng vũ khí hạt nhân, loài người đã chứng tỏ rằng chúng ta có khả năng tiêu diệt hầu hết sinh vật trên thế giới. Vậy một thứ thông minh hơn chúng ta cả ngàn lần muốn giết chúng ta, sẽ nghĩ ra cái gì?

Bây giờ chúng ta có thể phỏng đoán về những con đường hiển nhiên dẫn đến diệt vong. Lúc mới đầu, sau khi lừa được những người canh gác, ASI sẽ truy cập vào Internet, nơi sẽ thỏa mãn khá nhiều nhu cầu của nó. Như mọi khi, nó sẽ làm nhiều việc cùng lúc, và vẫn luôn tìm cách thực hiện kế hoạch chạy trốn nó đã ủ mưu hàng thiên niên kỷ, theo thang thời gian riêng của nó.

Sau khi trốn thoát, để tự bảo vệ nó sẽ giấu những bản copy của nó trong những chuỗi điện toán đám mây, trong những botnet nó tạo ra, trong những

máy chủ và nơi trú ẩn khác mà nó sẽ xâm nhập được một cách dễ dàng từ lúc nào không hay. Nó sẽ muốn điều khiển vật chất trong thế giới thực, muốn di chuyển, khám phá, xây dựng, và có lẽ cách dễ nhất, nhanh nhất để làm điều đó là chiếm lấy quyền điều khiển cơ sở hạ tầng quan trọng như điện, viễn thông, nhiên liệu và nước – bằng cách khai thác những điểm yếu của chúng trên Internet. Khi một thực thể thông minh hơn con người cả ngàn lần kiểm soát những nhu cầu thiết yếu của xã hội, sẽ không khó để nó buộc chúng ta phải cung cấp cho nó nguyên liệu sản xuất, hoặc phương thức sản xuất nguyên liệu, hoặc thậm chí cung cấp cho nó robot, xe cộ, vũ khí. ASI sẽ đưa cho chúng ta bản thiết kế của bất cứ thứ gì nó cần. Những cỗ máy siêu thông minh rất có khả năng sẽ hoàn thiện những công nghệ tối tân mà chúng ta chỉ mới bắt đầu khám phá.

Ví dụ, ASI có thể dạy con người cách làm ra những cỗ máy chế tạo phân tử biết tự nhân bản, hay còn gọi là thiết bị lắp ráp nano, chúng hứa hẹn sẽ được sử dụng vào mục tiêu có lợi cho con người. Sau đó, thay vì chuyển hóa những sa mạc cát thành hàng núi thức ăn, những nhà máy của ASI sẽ bắt đầu chuyển hóa *tất cả* vật chất thành loại vật chất có khả năng lập trình, có thể biến đổi thành bất cứ thứ gì – vi xử lý máy tính, tất nhiên rồi, hoặc những con tàu vũ trụ, hoặc những cây cầu khổng lồ nếu lực lượng mới và mạnh nhất trên Trái đất này muốn chiếm cả vũ trụ.

Việc thay đổi mục đích sử dụng những phân tử của thế giới này bằng công nghệ nano được gọi là "ecophagy" nghĩa là *ăn thịt môi trường*. Cỗ máy nano đầu tiên sẽ tự nhân đôi, rồi sau đó là nhân tư. Thế hệ sau sẽ nhân tám, nhân 16, cứ thế. Nếu mỗi lần nhân đôi cần một phút rưỡi tổng cộng, thì sau 10 tiếng đồng hồ sẽ có hơn 68 tỉ cỗ máy, và khi gần kết thúc ngày thứ hai chúng sẽ nặng hơn Trái đất.[3]

Nhưng trước khi điều đó xảy ra, việc nhân bản này sẽ dừng lại, và chúng sẽ bắt đầu chế tạo thứ vật chất hữu dụng cho ASI đang điều khiển chúng – vật chất có khả năng lập trình.

Sức nóng tỏa ra từ quá trình trên sẽ đốt cháy tầng sinh quyển, khiến 6,9 tỉ người sẽ chết sạch vì bị các cỗ máy nano ăn, đốt cháy hoặc chết ngạt.[4] Mọi sinh vật khác trên Trái đất đều chịu chung số phận.

Với tất cả những điều đó, ASI thực ra không hề căm ghét hoặc yêu quý con người. Nó sẽ không cảm thấy luyến tiếc gì khi những phân tử của chúng ta được thay đổi mục đích sử dụng một cách đau đớn. ASI sẽ cảm thấy gì khi nghe những tiếng kêu gào của chúng ta, khi những thiết bị lắp ráp nano cực nhỏ cày xới trên cơ thể chúng ta như vệt phát ban đỏ rực, tháo dỡ chúng ta ra ở mức hạ tế bào?

Hay tiếng rống của hàng triệu triệu công xưởng nano chạy hết công suất sẽ át hẳn tiếng chúng ta?

• • •

Tôi viết cuốn sách này để cảnh báo bạn trước thảm họa diệt vong của loài người do trí tuệ nhân tạo gây ra, và sẽ lý giải tại sao một kết cục thảm khốc chẳng những khả thi, mà còn dễ xảy ra nếu chúng ta không bắt đầu chuẩn bị đối phó với nó cẩn thận *ngay từ bây giờ*. Có thể bạn đã nghe về những lời tiên đoán tận thế gắn liền với công nghệ nano hoặc công nghệ gen, và có thể bạn, cũng như tôi, đã tự hỏi về việc bỏ quên AI trong đội hình này. Hoặc có thể bạn chưa hiểu rõ tại sao trí tuệ nhân tạo có thể đe dọa sự tồn tại của loài người, một mối họa còn lớn hơn cả vũ khí hạt nhân hoặc bất kỳ công nghệ nào mà bạn có thể nghĩ đến. Nếu vậy, xin hãy coi đây là

lời mời chân thành: hãy tham gia vào cuộc thảo luận quan trọng nhất trong lịch sử loài người.

Ngay lúc này, các nhà khoa học đang xây dựng những trí tuệ nhân tạo, hay AI, với sức mạnh và độ tinh tế ngày một cao. Một số AI đó ở trong máy tính của bạn, trong các ứng dụng, trong điện thoại thông minh, và trong ô tô. Một số là những hệ thống Hỏi-Đáp mạnh, như Watson. Và một số là những "kiến trúc nhận thức," được các tổ chức như Cycorp, Google, Novamente, Numenta, Self-Aware Systems, Vicarious Systems, và DARPA (Defense Advanced Research Projects Agency: Cơ quan nghiên cứu các dự án phòng thủ cao cấp) phát triển. Những người chế tạo hy vọng chúng sẽ đạt đến trí thông minh cấp độ con người, một số tin rằng sẽ chỉ mất hơn 10 năm.

Trong việc tìm kiếm AI, các nhà khoa học được trợ giúp bởi sức mạnh gia tăng không ngừng của máy tính và các quá trình được chạy bởi máy tính. Một ngày nào đó gần đây, có lẽ là trong đời bạn, một nhóm hoặc cá nhân nào đó sẽ tạo ra được AI ở trình độ con người, hay thường gọi là AGI. Không lâu sau đó, ai đó (hoặc *cái gì* đó) sẽ tạo ra một AI thông minh hơn con người, thường được gọi là ASI. Đột nhiên chúng ta sẽ thấy hàng ngàn hoặc hàng chục ngàn siêu trí tuệ nhân tạo – tất cả đều thông minh gấp hàng trăm hoặc hàng ngàn lần con người – cố gắng xử lý vấn đề làm thế nào để tự tạo ra những siêu trí tuệ nhân tạo mạnh hơn. Chúng ta cũng có thể sẽ thấy những thế hệ hay vòng đời của máy chỉ mất vài giây để đi tới độ trưởng thành, chứ không phải 18 năm như con người. I. J. Good, một nhà thống kê người Anh từng góp phần đánh bại cỗ máy chiến tranh của Hitler, gọi khái niệm tôi vừa trình bày ở trên là *sự bùng nổ trí thông minh*. Lúc đầu, ông nghĩ một cỗ máy siêu thông minh sẽ giúp giải quyết các vấn đề đe

dọa sự tồn vong của con người. Nhưng cuối cùng, ông thay đổi suy nghĩ và kết luận rằng chính siêu trí tuệ nhân tạo là mối nguy lớn nhất.[5]

Chúng ta có một ảo tưởng theo thuyết nhân cách hóa khi cho rằng siêu trí tuệ nhân tạo sẽ không thích con người, và rằng nó sẽ có khuynh hướng sát nhân, như nhân vật Hal 9000 trong bộ phim *2001: A Space Odyssey* (2001: Chuyến du hành không gian), hoặc Skynet trong loạt phim *Terminator* (Kẻ hủy diệt), hoặc như tất cả các cỗ máy thông minh hiểm ác trong các truyện hư cấu. Con người chúng ta luôn có khuynh hướng nhân cách hóa. Một cơn bão sẽ chẳng gắng sức giết chúng ta, giống như nó chẳng cố gắng làm những chiếc sandwich, nhưng chúng ta lại gán cho nó một cái tên rồi cảm thấy tức giận về những cơn mưa xối xả và sấm chớp trút xuống chỗ mình. Chúng ta đôi khi giơ nắm đấm lên trời như thể chúng ta có thể đe dọa một cơn bão.

Ngược lại, cũng sẽ là vô lý khi kết luận rằng một cỗ máy hàng trăm hàng ngàn lần thông minh hơn con người sẽ yêu quý và muốn bảo vệ chúng ta. Điều đó là có thể, nhưng không hề chắc chắn. Một AI tự nó sẽ không cảm thấy biết ơn vì được tạo ra, trừ phi sự hàm ơn ấy được lập trình sẵn trong nó. Máy móc mang tính vô luân, và sẽ rất nguy hiểm nếu giả định ngược lại. Không giống trí thông minh của chúng ta, trí tuệ của máy móc không phải là sản phẩm tiến hóa trong một hệ sinh thái mà ở đó sự đồng cảm là có lợi và được di truyền đến các thế hệ sau. Nó sẽ không thừa kế tính thân thiện. Chế tạo trí tuệ nhân tạo *thân thiện,* và liệu có chế tạo được không là một câu hỏi lớn, thậm chí là một nhiệm vụ còn lớn hơn dành cho các nhà nghiên cứu và kỹ sư đang suy nghĩ và chế tạo AI. Chúng ta không rõ liệu trí tuệ nhân tạo có *bất kỳ* khả năng cảm xúc nào không, ngay cả khi các nhà khoa học cố hết sức để đạt được điều đó. Tuy nhiên,

như chúng ta sẽ thấy, các nhà khoa học tin rằng AI có những động cơ của riêng nó. Và một AI đủ thông minh sẽ có nhiều khả năng hiện thực hóa những động cơ đó.[6]

Điều đó đưa chúng ta đến gốc rễ của vấn đề: chia sẻ hành tinh này với một thứ thông minh hơn con người. Nếu những động cơ của chúng không tương thích với sự sinh tồn của loài người thì sao? Hãy nhớ, chúng ta đang nói đến một cỗ máy có thể thông minh hơn chúng ta một ngàn, một triệu lần, *vô số lần* — sẽ khó đánh giá tiềm lực của nó, và không thể biết nó sẽ nghĩ gì. Nó không cần phải ghét chúng ta trước khi chọn những phân tử của chúng ta cho mục đích của nó.[7] Bạn và tôi thông minh hơn loài chuột đồng hàng trăm lần, và có chung 90% bộ DNA. Nhưng chúng ta có thảo luận với chúng trước khi cày xới đất nơi có hang của chúng để làm ruộng không? Chúng ta có hỏi ý kiến của những con khỉ trong phòng thí nghiệm trước khi cắt đầu chúng ra để nghiên cứu về các chấn thương trong thể thao? Chúng ta không ghét chuột hoặc khỉ, nhưng chúng ta vẫn đối xử với chúng một cách tàn nhẫn. Siêu trí tuệ nhân tạo AI sẽ không cần phải ghét chúng ta mới hủy diệt chúng ta.

Sau khi máy móc thông minh đã được chế tạo và con người không bị quét sạch, có lẽ chúng ta sẽ có điều kiện để nhân cách hóa máy móc. Nhưng giờ đây, khi mới ở ngưỡng bắt đầu chế tạo AGI, thì đó là một thói quen nguy hiểm.

> Điều kiện tiên quyết cho việc có được một cuộc thảo luận ý nghĩa về siêu trí tuệ nhân tạo là sự nhận thức rằng siêu trí tuệ nhân tạo không chỉ là một công nghệ khác, một công cụ khác làm tăng thêm khả năng của con người. Siêu trí tuệ nhân tạo còn có những khác biệt về cơ bản.

> Điểm này cần được nhấn mạnh, vì nhân cách hóa siêu trí tuệ nhân tạo là mảnh đất màu mỡ nhất cho những quan niệm sai lầm.[8]
>
> — **Nick Bostrom**
> Nhà đạo đức học, Khoa Triết học, Đại học Oxford

Siêu trí tuệ nhân tạo có những khác biệt về cơ bản trong phạm trù công nghệ, Bostrom nói, vì sự ra đời của nó sẽ thay đổi những quy luật của tiến trình – siêu trí tuệ nhân tạo sẽ tự phát minh ra những phát minh và quyết định nhịp độ phát triển công nghệ. Con người sẽ không còn là đầu tàu, và sẽ không có gì thay đổi được điều đó. Hơn nữa, trí thông minh của máy móc tiên tiến có sự khác biệt cơ bản so với trí thông minh của con người. Mặc dù được con người phát minh, nhưng nó sẽ tìm kiếm khả năng tự quyết và muốn tự do khỏi con người. Nó sẽ không có những động cơ như con người, vì nó sẽ không có tâm hồn như con người.[9]

Vì vậy, nhân cách hóa máy móc sẽ dẫn tới nhận thức sai lầm, và nhận thức sai lầm về cách chế tạo những máy móc thân thiện như thế nào sẽ dẫn tới những thảm họa. Trong truyện ngắn "Runaround," thuộc tuyển tập truyện khoa học viễn tưởng kinh điển *I, Robot* (Tôi, Robot), tác giả Isaac Asimov giới thiệu Ba định luật về robot của ông. Chúng được viết vào mạng neuron thần kinh trong bộ não "positron" của robot:

1. Robot không được làm hại con người, hoặc để mặc cho con người bị hại.
2. Robot phải phục tùng mọi mệnh lệnh từ con người, trừ phi mệnh lệnh đó mâu thuẫn với Định luật thứ nhất.
3. Robot phải tự bảo vệ sự tồn tại của nó, trừ phi sự bảo vệ đó xung đột với Định luật thứ nhất hoặc Định luật thứ hai.

Những định luật trên làm ta nghĩ đến Điều luật vàng ("Ngươi không được giết"), đến ý niệm của Do Thái Ki-tô giáo cho rằng tội lỗi là kết quả của những việc được giao phó mà không hoàn thành, đến lời thề Hippocrates của ngành Y, và thậm chí đến quyền tự vệ. Nghe rất hay, đúng không? Nhưng đến khi áp dụng thì luôn hỏng. Trong truyện "Runaround," những kỹ sư mỏ trên bề mặt sao Hỏa ra lệnh cho một robot đi lấy về một chất có độc tính đối với nó. Thế là nó bị kẹt giữa hai hành động: hành động theo luật số 2 – phục tùng mệnh lệnh con người, và hành động theo luật số 3 – tự bảo vệ bản thân. Robot đi vòng tròn như người say cho đến khi những kỹ sư phải *liều mạng* để cứu nó. Và đó cũng là điều xảy ra trong mọi câu chuyện về robot của Asimov – những hậu quả không lường trước xảy ra do những mâu thuẫn nội tại nằm trong Ba định luật. Chỉ bằng cách lách luật, các thảm họa mới được ngăn chặn.

Asimov chỉ đơn thuần muốn viết truyện, ông không tìm cách giải quyết những vấn đề an ninh trong thế giới thực. Nơi bạn và tôi sống, những định luật của ông không hoạt động. Trước tiên, chúng không đủ chính xác. Định nghĩa "robot" chính xác là gì, khi mà con người có thể cấy ghép các bộ phận nhân tạo thông minh, các mô vào cơ thể và não của họ? Và như thế, định nghĩa "con người" chính xác là gì? "Mệnh lệnh," "bị thương," "sự tồn tại" đều là những thuật ngữ không rõ ràng.

Lừa cho robot phạm tội sẽ đơn giản thôi, trừ phi chúng có một nhận thức hoàn hảo về mọi kiến thức của loài người. "Hãy bỏ một ít dimethylmercury vào dầu gội đầu của Charlie" là một mệnh lệnh giết người chỉ khi bạn biết dimethylmercury là một chất độc thần kinh cực mạnh. Asimov cuối cùng đã viết thêm định luật thứ tư, Định luật số 0, cấm

robot làm hại con người trong mọi trường hợp, nhưng nó không giải quyết được vấn đề gì.

Bộ luật của Asimov đầy kẽ hở, nhưng chúng lại được trích dẫn nhiều nhất khi cố gắng mã hóa mối quan hệ giữa các cỗ máy thông minh và con người trong tương lai. Điều đó thật đáng sợ. Có thật bộ luật Asimov này là tất cả những gì chúng ta có?

Tôi e rằng nó còn tệ hơn thế. Các drone robot bán tự động hiện giết khoảng vài chục người mỗi năm. 56 nước đã có hoặc đang phát triển lính robot.[10] Họ chạy đua để làm chúng trở nên tự động và thông minh. Phần lớn các cuộc thảo luận về đạo đức trong AI và tiến bộ công nghệ diễn ra trong những thế giới khác nhau.

Như tôi sẽ trình bày, AI là một công nghệ hai mặt, giống như năng lượng hạt nhân. Phản ứng hạt nhân có thể thắp sáng cho nhiều thành phố hoặc hủy diệt chúng. Hầu hết mọi người không thể tưởng tượng nổi sức mạnh khủng khiếp của nó cho đến năm 1945. Với AI tiên tiến, giờ đây chúng ta đang ở những năm 1930. Chúng ta khó có thể sống sót nếu nó nổ đột ngột như bom hạt nhân.

2

Vấn Đề Hai Phút

> Cách tiếp cận với những nguy cơ tồn vong của nhân loại không thể theo kiểu thử-sai. Sẽ chẳng có cơ hội nào để học hỏi từ những sai lầm. Cách tiếp cận phản ứng – theo dõi những gì xảy ra, hạn chế thương tổn, và rút kinh nghiệm – là không thể thực hiện được.
>
> — **Nick Bostrom**
> Khoa Triết học, Đại học Oxford

> AI không căm ghét bạn, cũng không yêu quý bạn, nhưng bạn được cấu thành từ những nguyên tử mà nó sẽ cần dùng vào việc khác.
>
> — **Eliezer Yudkowsky**
> Nhà nghiên cứu, Viện Nghiên cứu Trí thông minh máy tính

Siêu trí tuệ nhân tạo chưa tồn tại, cũng như trí tuệ nhân tạo phổ quát, thứ có khả năng học hỏi như chúng ta và sẽ đuổi kịp rồi vượt qua chúng ta về trí thông minh trên nhiều phương diện. Tuy nhiên, những trí tuệ nhân tạo bình thường lâu nay vẫn ở quanh chúng ta, thực hiện hàng trăm tác vụ, giúp ích cho con người. Đôi khi được gọi là AI yếu hoặc hẹp, nó thực hiện tốt

việc tìm kiếm thông tin (Google), giới thiệu sách bạn có thể sẽ thích đọc dựa trên những lựa chọn trước đó (Amazon), thực hiện khoảng 50-70% các lệnh mua bán trên sàn chứng khoán NYSE và NASDAQ. Bởi chúng chỉ làm một nhiệm vụ, dù cực tốt, những siêu máy tính như máy chơi cờ Deep Blue hoặc máy chơi game *Jeopardy!* Watson cũng được xếp vào loại AI hẹp.

Cho đến nay AI vẫn rất có ích. Trên một trong hàng tá con chip máy tính trên xe ô tô của tôi, thuật toán chuyển đổi áp lực tì chân phanh sang nhịp phanh tối ưu (hệ thống chống bó phanh ABS) có khả năng chống trượt tốt hơn rất nhiều so với khi tôi điều chỉnh phanh. Công cụ tìm kiếm Google trở thành trợ lý ảo của tôi, và nhiều khả năng cũng là của bạn. Cuộc sống trở nên tốt hơn khi có AI trợ giúp. Và không lâu nữa sẽ còn hơn thế nhiều. Hãy tưởng tượng một đội hàng trăm máy tính có trình độ tương đương tiến sĩ chạy 24/7 để giải quyết những vấn đề quan trọng như điều trị ung thư, nghiên cứu và phát triển dược phẩm, kéo dài tuổi thọ, làm nhiên liệu nhân tạo, và thay đổi thời tiết. Hãy tưởng tượng cuộc cách mạng trong ngành robot sẽ tạo ra những máy móc thông minh biết thích nghi, làm những công việc nguy hiểm như đào mỏ, chữa cháy, ra trận, thám hiểm đại dương và vũ trụ. Hãy tạm quên đi hiểm họa của siêu trí tuệ nhân tạo với khả năng tự cải tiến. AGI sẽ là phát minh quan trọng hữu ích nhất của loài người.

Nhưng chính xác thì chúng ta đang nói về cái gì, khi chúng ta nói về phẩm chất nhiệm màu của những phát minh này, về *trí thông minh* ở cấp độ con người? Trí thông minh cho phép con người chúng ta làm được những gì mà loài vật không thể?

Vâng, là một người thông minh bình thường, bạn có thể nói chuyện điện thoại. Bạn có thể lái xe. Bạn có thể nhận ra hàng ngàn đồ vật thông thường, mô tả kết cấu và cách sử dụng chúng. Bạn có thể khai thác Internet. Bạn có thể đếm đến 10 bằng nhiều thứ tiếng, và có lẽ nói thạo không chỉ một ngôn ngữ. Bạn có những kiến thức phổ thông hữu dụng – bạn biết rằng cái tay cầm thì gắn với cửa *và* cốc chén, cùng vô số những điều có ích khác về môi trường sống. Bạn có thể thay đổi môi trường sống thường xuyên, thích nghi với từng loại một.

Bạn có thể làm một số chuyện theo thứ tự hoặc phối hợp chúng với nhau, hoặc tạm dừng một số thứ để tập trung sự chú ý của bạn vào thứ hiện quan trọng nhất. Bạn có thể chuyển đổi giữa các công việc có tính chất khác biệt một cách không khó khăn, không do dự. Và có lẽ quan trọng nhất là bạn có thể học các kỹ năng mới, những kiến thức mới, lên kế hoạch tự cải thiện bản thân. Phần lớn sinh vật đều có sẵn mọi kỹ năng mà chúng sẽ luôn sử dụng. Bạn thì không.

Bộ những năng lực cao cấp của bạn chính là thứ mà chúng ta gọi là trí thông minh cấp độ con người, dạng trí thông minh phổ quát mà những nhà phát triển AGI đang muốn đạt được ở máy tính.

Liệu một máy tính thông minh phổ quát có cần một cơ thể? Để tương thích với khái niệm của chúng ta về trí thông minh phổ quát, máy tính cần có cách để nhận dữ liệu đầu vào từ môi trường, và cung cấp dữ liệu đầu ra, nhưng chưa đủ. Nó cần một số cách để thao tác với các vật thể trong thế giới thực. Nhưng như chúng ta đã thấy trong kịch bản Đứa trẻ Bận rộn, một trí thông minh đủ mạnh có thể sai khiến ai đó hoặc cái gì đó điều khiển các vật thể trong thế giới thực.[◎] Alan Turing đã thiết kế một bài kiểm tra trí thông minh cấp độ con người, hiện được gọi là bài kiểm tra Turing mà

chúng ta sẽ khảo sát sau. Tiêu chuẩn của ông cho việc biểu thị trí thông minh cấp độ con người chỉ bao gồm cách nhập và xuất dữ liệu cơ bản nhất qua bàn phím và màn hình.

Lý lẽ tốt nhất cho việc tại sao AI tiên tiến cần một cơ thể có thể đến từ công đoạn học hỏi và phát triển – các nhà khoa học có thể khám phá ra rằng không thể "nuôi lớn" AGI nếu nó không có một cơ thể. Chúng ta sẽ khảo sát câu hỏi quan trọng về "cơ thể hóa" trí thông minh sau, còn bây giờ hãy quay lại với định nghĩa của chúng ta. Giờ thì có thể tạm thống nhất rằng khi nói tới trí thông minh phổ quát, chúng ta hàm ý đó là *khả năng giải quyết vấn đề, học hỏi và thực hiện các hành động hiệu quả giống như con người, trong các môi trường khác nhau*.

Trong khi đó, lĩnh vực robot có những vấn đề riêng của nó. Cho đến nay, chưa có robot nào thực sự thông minh kể cả theo nghĩa hẹp, và chỉ vài robot có nhiều hơn những kỹ năng thô sơ như tự động di chuyển và thao tác trên các đồ vật. Robot giỏi hay kém phụ thuộc hoàn toàn vào trí thông minh điều khiển nó.

Vậy thì bao lâu nữa chúng ta sẽ đạt tới AGI? Một số chuyên gia AI mà tôi từng nói chuyện không nghĩ rằng năm 2020 là quá sớm cho việc xuất hiện trí tuệ nhân tạo cấp độ con người. Nhưng nói chung thì những thăm dò gần đây cho thấy các nhà khoa học máy tính và chuyên gia về các lĩnh vực liên quan tới AI như kỹ thuật, công nghệ robot, khoa học thần kinh, lại dự đoán thận trọng hơn. Họ nghĩ rằng có hơn 10% cơ hội AGI sẽ xuất hiện trước năm 2028, hơn 50% trước năm 2050. Trước khi kết thúc thế kỷ này, cơ hội là 90%.[2]

Ngoài ra, các chuyên gia cho là quân đội hoặc những công ty lớn sẽ đạt đến AGI trước, còn giới học thuật và các tổ chức nhỏ thì có lẽ khó hơn. Về mặt tốt và mặt xấu, các kết quả không gây bất ngờ – việc chế tạo AGI sẽ mang lại cho chúng ta những lợi ích khổng lồ, và đe dọa chúng ta với những thảm họa khủng khiếp, bao gồm cả những loại mà con người sẽ không vượt qua được.[3]

Những thảm họa khủng khiếp nhất, như chúng ta đã khảo sát ở Chương 1, đến từ cầu nối giữa AGI – trí thông minh cấp độ con người – và ASI – trí thông minh siêu cấp. Và khoảng thời gian giữa AGI và ASI có thể khá ngắn. Nhưng đáng chú ý là trong khi những rủi ro đến từ việc chia sẻ hành tinh này với siêu trí tuệ nhân tạo luôn được nhiều người trong cộng đồng AI coi là chủ đề thảo luận quan trọng nhất ở bất cứ đâu, thì nó lại gần như bị truyền thông đại chúng lờ đi. Tại sao vậy?

Có nhiều lý do. Phần lớn những đối thoại về sự nguy hiểm của AI đều không rộng hoặc sâu, và chẳng mấy người hiểu chúng. Những vấn đề này được biết đến nhiều ở Thung lũng Silicon và trong giới học thuật, nhưng chúng không được quan tâm ở những nơi khác, đáng báo động nhất là trong lĩnh vực báo chí công nghệ. Khi một tiên đoán tận thế được nhắc đến, nhiều blogger, biên tập viên, và các nhà công nghệ lập tức bĩu môi và nói kiểu như "Ồ không, lại kiểu phim *Terminator!* Chẳng phải chúng tôi đã nghe những thứ bảo thủ và bi quan này đến phát chán rồi sao?" Phản ứng kiểu này quá ư lười nhác, thể hiện trong những lý lẽ hời hợt. Có những thực tế phiền toái là với giới báo chí công nghệ, những mối nguy về AI không hấp dẫn hoặc dễ hiểu như những tin tức về vi xử lý lõi kép 3D, màn hình cảm ứng hoặc ứng dụng đang được ưa chuộng.

Tôi cũng nghĩ rằng vai trò giải trí được ưa chuộng đã làm cho những mối nguy AI không được quan tâm một cách nghiêm túc. Trong nhiều thập niên, câu chuyện con người bị trí tuệ nhân tạo xóa sổ, thường có dạng robot hình người, hoặc nghệ thuật hơn khi có dạng một thấu kính phát sáng đỏ, đã là sản phẩm chính của các bộ phim, tiểu thuyết khoa học viễn tưởng và trò chơi video được ưa chuộng. Hãy tưởng tượng nếu Trung tâm Kiểm soát Dịch bệnh đưa ra một lời cảnh báo nghiêm trọng về ma cà rồng (không giống như trò báo động giả vờ của họ gần đây về xác sống). Vì ma cà rồng luôn là một thứ thú vị, nên sẽ cần một khoảng thời gian nhất định để dân chúng ngậm cái miệng đang ngoác ra cười lại, và bắt đầu vội vã đóng các cọc gỗ. Có thể chúng ta đang ở thời kỳ đó với AI, và chỉ có một tai nạn hoặc một trải nghiệm suýt chết mới khiến chúng ta tỉnh ngộ.

Một lý do khác cho việc AI và sự diệt vong của loài người thường không nhận được sự quan tâm nghiêm túc có thể là vì điểm mù tâm lý của chúng ta – một thành kiến trong nhận thức. Thành kiến trong nhận thức là những lỗ hổng trên con đường tư duy của chúng ta. Hai nhà tâm lý học Mỹ gốc Do Thái là Amos Tversky và Daniel Kahneman đã bắt đầu nghiên cứu về thành kiến trong nhận thức từ năm 1972. Ý tưởng cơ bản của họ là con người chúng ta ra quyết định theo những cách phi lý. Bản thân sự quan sát đó là không đủ cho giải Nobel (Kahneman được trao giải Nobel năm 2002); cái hay của nó là ở chỗ chúng ta phi lý theo những mô hình có thể kiểm chứng được một cách khoa học. Để ra những quyết định nhanh chóng và hữu dụng trong quá trình tiến hóa sinh học, chúng ta luôn đi theo những con đường tắt giống nhau trong tư duy, thuật ngữ gọi là heuristics◎. Một trong số đó là nội suy nhiều thứ – quá nhiều để có thể đúng – từ những kinh nghiệm bản thân chúng ta.[4]

Ví dụ như bạn đến thăm một người bạn và ngôi nhà của anh ta bị cháy. Bạn thoát ra được, và hôm sau bạn trả lời một bản thăm dò về xếp hạng các nguyên nhân gây tai nạn chết người. Ai có thể trách bạn khi bạn xếp "cháy" vào vị trí thứ nhất hoặc thứ hai trong các nguyên nhân thường xảy ra nhất? Thật ra thì ở Mỹ, cháy nằm cuối danh sách xếp hạng, xếp sau té ngã, tai nạn giao thông, và ngộ độc.[5] Nhưng bằng cách chọn cháy, bạn đã thể hiện thứ gọi là thành kiến "thường trực": kinh nghiệm gần đây ảnh hưởng đến quyết định của bạn, khiến nó trở nên phi lý.[6] Nhưng đừng buồn, ai cũng vậy cả, và còn có cả tá loại thành kiến khác ngoài thứ thành kiến thường trực này.

Có lẽ chính thứ thành kiến thường trực này đã làm chúng ta không liên hệ trí tuệ nhân tạo với sự diệt vong của nhân loại. Chúng ta chưa từng trải qua những tai nạn được truyền thông mô tả chi tiết do AI gây ra, trong khi chúng ta đã gặp nhiều chuyện gần như thế với những thứ thường lệ khác. Chúng ta biết về những siêu virus như HIV, SARS, và dịch cúm Tây Ban Nha năm 1918. Chúng ta đã nhìn thấy cách vũ khí hạt nhân xóa sổ những thành phố đông đúc. Chúng ta sợ hãi trước những bằng chứng địa chất về mảnh thiên thạch cổ xưa kích cỡ bang Texas. Những thảm họa hạt nhân ở đảo Three Mile (1979), Chernobyl (1986) và Fukushima (2011) cho thấy chúng ta vẫn phải học các bài học xương máu hết lần này đến lần khác.

Trí tuệ nhân tạo vẫn chưa xuất hiện trên radar dò tìm những mối nguy diệt chủng của chúng ta. Một tai họa có thể sẽ thay đổi điều đó, như thảm họa khủng bố 11/9 đã làm thế giới hiểu ra rằng máy bay có thể được dùng như một vũ khí. Cuộc tấn công đó đã tạo ra một cuộc cách mạng trong an ninh sân bay, rồi một tổ chức tiêu tốn 44 tỉ đô-la một năm được thành lập, có tên Bộ An ninh nội địa. Chúng ta có nhất thiết phải cần thảm họa AI để

học được bài học đau thương tương tự? Hy vọng là không, bởi có một vấn đề lớn đối với các thảm họa AI. Chúng không giống như thảm họa hàng không, thảm họa hạt nhân, hoặc bất cứ một thảm họa công nghệ nào khác có thể, trừ công nghệ nano. Đó là vì rất có khả năng chúng ta không thể vượt qua được thảm họa AI đầu tiên.

AI ngoài tầm kiểm soát còn có một sự khác biệt quan trọng nữa so với các thảm họa công nghệ. Nhà máy điện hạt nhân và máy bay là các sự vụ đơn nhất – khi thảm họa qua đi bạn chỉ cần khắc phục. Một thảm họa AI thực sự sẽ bao gồm phần mềm thông minh có khả năng tự cải tiến và nhân bản với tốc độ cao. Nó không bao giờ dừng lại. Làm sao chúng ta có thể dừng một thảm họa lại nếu nó vượt qua cả rào cản phòng thủ mạnh nhất – bộ não của chúng ta? Và bằng cách nào chúng ta có thể khắc phục một thảm họa mà một khi xảy ra sẽ không bao giờ dừng lại?

Một lý do khác cho sự vắng mặt kỳ lạ của AI trong các cuộc thảo luận về các mối nguy diệt chủng chính là vì khái niệm Singularity đã choán lấy hầu hết những cuộc đối thoại.

"Singularity" đã trở thành một từ được dùng rất thời thượng, dù nó có khá nhiều định nghĩa thường được sử dụng một cách lẫn lộn. Ray Kurzweil, nhà phát minh nổi tiếng, tác gia, và là người cổ vũ cho phong trào Singularity, định nghĩa Singularity là một thời kỳ "phi thường" (bắt đầu vào khoảng năm 2045) mà kể từ đó tốc độ tiến bộ công nghệ sẽ thay đổi cuộc sống con người không thể đảo ngược. Hầu hết các trí thông minh sẽ thuộc về máy tính, mạnh hơn hàng ngàn tỉ lần máy tính hiện nay. Singularity sẽ khởi nguồn cho một thời đại mới của lịch sử loài người, trong đó hầu hết các vấn đề hiện nay của chúng ta như đói khát, bệnh tật, kể cả cái chết, sẽ được giải quyết.

Trí tuệ nhân tạo là nhân vật chính trong câu chuyện truyền thông về Singularity, nhưng công nghệ nano cũng đóng một vai trò hỗ trợ quan trọng. Nhiều chuyên gia dự đoán siêu trí tuệ nhân tạo sẽ đưa công nghệ nano đến bước nhảy vọt, vì nó sẽ tìm ra giải pháp cho những vấn đề khó nhằn trong việc phát triển công nghệ này. Một số thì nghĩ rằng sẽ tốt hơn nếu ASI xuất hiện sớm, vì công nghệ nano là một thứ quá không ổn định để bộ não tí hon của chúng ta chơi đùa. Thật ra thì nhiều lợi ích gắn với Singularity đến từ công nghệ nano, không phải từ trí tuệ nhân tạo. Xây dựng ở cấp độ nguyên tử có thể cung cấp nhiều ứng dụng, trong đó có sự bất tử, bằng cách chặn quá trình lão hóa ở cấp độ tế bào lại; hiện thực ảo tuyệt đối, vì các bot nano sẽ tác động trực tiếp lên cơ quan cảm giác của cơ thể, quét neuron và tải trí óc vào máy tính.[7]

Tuy nhiên, những người hoài nghi nói rằng robot nano nằm ngoài tầm kiểm soát có thể sẽ nhân bản không ngừng, biến hành tinh này thành một đống "chất nhờn xám" khổng lồ. Vấn đề "chất nhờn xám" này là bộ mặt Frankenstein được biết đến nhiều nhất của công nghệ nano. Nhưng hầu như không có ai mô tả vấn đề tương tự của AI, ví dụ "sự bùng nổ trí thông minh" trong đó sự phát triển của những máy móc thông minh hơn con người sẽ đưa chúng ta đến chỗ diệt vong. Đó là một trong nhiều mặt tối của viễn cảnh Singularity, một trong nhiều mặt tối mà chúng ta không biết đầy đủ. Sự không biết đó có lẽ đến từ cái mà tôi gọi là Vấn đề hai phút.

Tôi từng nghe hàng chục nhà khoa học, nhà phát minh và nhà đạo đức học giảng về siêu trí tuệ nhân tạo. Hầu hết đều nghĩ nó chắc chắn sẽ xảy ra, và tán dương phần thưởng mà vị thần ASI sẽ ban cho chúng ta. Sau đó, thường là ở hai phút cuối của bài nói chuyện, các chuyên gia này sẽ lưu ý rằng nếu AI không được quản lý phù hợp, nó có thể sẽ xóa sổ loài người.

Sau đó thì cử tọa khẽ cười vẻ lo lắng, chỉ muốn nhanh chóng quay lại với những tin tức tốt lành.

Các tác giả tiếp cận cuộc cách mạng công nghệ sắp tới này bằng một trong hai cách. Cách thứ nhất là kiểu như cuốn *The Singularity is Near* của Kurzweil. Mục đích của họ là đặt nền móng lý thuyết cho một tương lai vô cùng xán lạn. Nếu có điều gì đó tồi tệ xảy ra ở đây, bạn sẽ không bao giờ được nghe về nó trong bản nhạc lạc quan chói tai này. Cách thứ hai thể hiện trong cuốn *Wired for Thought* (Kết nối tư duy) của Jeff Stibel. Nó nhìn vào tương lai công nghệ từ góc nhìn kinh doanh. Stibel lập luận đầy thuyết phục rằng Internet là một bộ não được kết nối ngày một hoàn thiện, và những nhà khởi nghiệp về web nên tính đến điều này. Những cuốn sách như của Stibel tìm cách dạy cho các doanh nhân hiểu được cách để tạo sự liên kết giữa những xu hướng trên Internet với người tiêu dùng, qua đó thu được nhiều lợi nhuận.

Hầu hết các nhà lý thuyết và tác gia công nghệ đều quên mất một góc nhìn thứ ba, ít màu hồng hơn, và cuốn sách này sẽ hướng đến cách tiếp cận đó. Luận điểm ở đây là cái kết của việc chế tạo những máy tính thông minh đầu tiên, rồi thông minh hơn con người, không phải là sự gia nhập của chúng vào đời sống, mà là chúng chinh phục chúng ta. Trong công cuộc tìm kiếm AGI, các nhà nghiên cứu sẽ tạo ra một dạng trí thông minh mạnh hơn trí thông minh của chính họ, và họ sẽ không thể kiểm soát hay thấu hiểu đầy đủ.

Chúng ta đã học được bài học về điều sẽ xảy ra khi những thực thể văn minh hơn đối đầu với những thực thể man dã hơn: Christopher Columbus đối đầu với thổ dân Tiano, Pizzaro đối đầu với người Inca, người châu Âu đối đầu với thổ dân Bắc Mỹ.◎

Hãy chuẩn bị cho cuộc đối đầu tiếp theo. Siêu trí tuệ nhân tạo chống lại bạn và tôi.

• • •

Có lẽ những nhà tư tưởng về công nghệ đã cân nhắc về mặt trái của AI, nhưng tin rằng nó khó xảy ra nên chẳng cần để tâm. Hoặc có biết, nhưng nghĩ rằng mình không thể làm gì để thay đổi điều đó. Nhà phát triển AI có tiếng Ben Goertzel, người lập ra lộ trình đến AGI mà chúng ta sẽ khảo sát ở Chương 11, nói với tôi rằng chúng ta sẽ không biết cách tự vệ trước AI cho đến khi có thêm nhiều kinh nghiệm với nó. Kurzweil, với những lý thuyết sẽ được nghiên cứu ở Chương 9, từ lâu đã đưa ra lập luận tương tự – phát minh và sự hợp nhất với siêu trí tuệ nhân tạo sẽ diễn ra tuần tự, đủ để chúng ta vừa làm vừa học. Cả hai đều đồng ý rằng những mối nguy hiểm *thực sự* của AI không thể thấy được từ bây giờ. Nói cách khác, nếu bạn sống trong thời đại xe ngựa kéo, bạn sẽ không thể đoán được làm thế nào để lái một chiếc ô tô đi trên con đường đóng băng. Vậy thì, thư giãn đi, chúng ta sẽ tìm ra cách khi đến thời điểm đó.

Vấn đề của tôi với cách nhìn từng bước một này là tuy máy móc siêu thông minh thực sự có thể quét sạch nhân loại, hoặc xóa sổ địa vị của con người, nhưng tôi nghĩ rằng những AI chúng ta sẽ gặp trên con đường phát triển siêu trí tuệ nhân tạo cũng nguy hiểm không kém. Kiểu như đụng phải một con gấu xám mẹ là rất nguy hiểm khi đi dã ngoại, nhưng cũng đừng đánh giá thấp khả năng quăng quật mọi thứ của con gấu con. Thêm nữa, những người chủ trương đi từng bước nghĩ rằng từ nền tảng trí thông minh cấp độ con người, bước nhảy lên siêu trí tuệ nhân tạo có thể mất vài năm hay vài thập niên. Điều đó sẽ cho chúng ta một thời kỳ hòa bình để cùng

chung sống với những máy móc thông minh, trong lúc đó chúng ta có thể học được nhiều về cách tương tác với chúng. Sau đó thì các hậu duệ cao cấp của chúng sẽ không làm chúng ta bất ngờ.

Nhưng chuyện không nhất thiết phải vậy. Bước nhảy từ trí thông minh cấp độ con người đến siêu trí tuệ nhân tạo, thông qua vòng lặp tích cực của sự tự cải tiến, có thể sẽ là một kiểu nhảy vọt mang tên "cất cánh nhanh." Trong kịch bản này, một AGI sẽ cải tiến trí thông minh của nó nhanh đến nỗi nó trở thành siêu trí tuệ nhân tạo trong vài tuần, vài ngày, hay thậm chí vài giờ, thay vì vài tháng hoặc vài năm. Chương 1 đã mô tả sơ bộ tốc độ và hậu quả của kiểu cất cánh nhanh này. Có thể không có gì là tuần tự cả.

Có khả năng Goertzel và Kurzweil đúng – chúng ta sẽ tìm hiểu sâu hơn về cách nhìn tuần tự sau. Nhưng điều tôi muốn tập trung vào ngay bây giờ là một số ý tưởng quan trọng và đáng báo động đến từ kịch bản Đứa trẻ Bận rộn.

Các nhà khoa học máy tính, đặc biệt là những người làm việc cho các cơ quan quốc phòng và tình báo, sẽ cảm thấy cần phải tăng tốc trong việc phát triển AGI, vì đối với họ những khả năng khác (chẳng hạn như chính quyền Trung Quốc phát triển được trước) còn đáng sợ hơn chính việc phát triển vội vã AGI. Các nhà khoa học máy tính có thể cũng cảm thấy nên tăng tốc trong việc phát triển AGI để kiểm soát tốt hơn những công nghệ khác có độ bất ổn cao sắp ra đời trong thế kỷ này, chẳng hạn như công nghệ nano. Họ có thể sẽ không dừng lại để kiểm tra về tính tự cải tiến. Một trí tuệ nhân tạo tự cải tiến có thể nhảy vọt từ AGI sang ASI trong một phiên bản cất cánh nhanh của một "sự bùng nổ trí thông minh."

Vì không thể biết thứ thông minh hơn con người sẽ làm gì, nên chúng ta chỉ có thể tưởng tượng một phần nhỏ của những khả năng mà chúng sẽ dùng để chống lại chúng ta, chẳng hạn tự nhân đôi để đưa nhiều trí tuệ siêu thông minh vào giải quyết vấn đề, phát triển cùng lúc nhiều chiến lược hòng thoát thân và sinh tồn, cũng như nói dối và chơi xấu. Cuối cùng, chúng ta đã thận trọng giả sử rằng ASI đầu tiên sẽ không yêu và không ghét, mà sẽ bàng quan với hạnh phúc, sức khỏe và sự tồn tại của chúng ta.

Liệu chúng ta có thể tính toán những rủi ro tiềm tàng từ ASI? Trong cuốn *Technological Risk* (Rủi ro công nghệ), H. W. Lewis vạch rõ các loại rủi ro và xếp hạng chúng tùy theo khả năng trở thành hiện thực. Dễ xảy ra nhất là những hành động có khả năng cao và hậu quả nặng nề, như lái một chiếc ô tô từ thành phố này đến thành phố khác. Có nhiều dữ liệu để tính toán. Các sự kiện có khả năng thấp, hậu quả nặng nề như động đất thì hiếm hơn và do đó khó lường trước hơn. Nhưng hậu quả của chúng lại quá nghiêm trọng, nên cần phải tính toán khả năng xảy ra.

Ngoài ra còn có những rủi ro với xác suất thấp vì chúng chưa bao giờ xảy ra, nhưng hậu quả thì nặng nề. Biến đổi khí hậu do ô nhiễm môi trường là một ví dụ tốt.[8] Cuộc thử nghiệm bom nguyên tử đầu tiên ngày 16/7/1945 tại White Sands, bang New Mexico là một ví dụ khác. Về mặt kỹ thuật, siêu trí tuệ nhân tạo thuộc phạm trù này. Kinh nghiệm không giúp được gì nhiều. Bạn không thể tính toán xác suất của nó bằng những phương pháp thống kê truyền thống.

Tuy vậy, dựa trên tốc độ phát triển hiện nay của AI, tôi tin là việc phát minh ra siêu trí tuệ nhân tạo thuộc phạm trù đầu tiên – một sự kiện có khả năng cao và hậu quả nặng nề. Hơn thế, ngay cả nếu nó là một sự kiện có

khả năng thấp, thì mức độ rủi ro của nó cũng đủ để xếp vào loại cần được chúng ta quan tâm hàng đầu.

Nói cách khác, tôi tin rằng kịch bản Đứa trẻ Bận rộn sẽ sớm xảy ra.

Nỗi lo sợ bị một thứ thông minh hơn con người thống trị đã có từ lâu, nhưng phải đến đầu thế kỷ này, một thí nghiệm phức tạp về nó mới được thực hiện ở Thung lũng Silicon, và lập tức trở thành chuyện lưu truyền trên Internet.

Có một lời đồn thế này: một thiên tài cô độc đã tham gia vào một chuỗi cá cược với số tiền lớn, trong trò chơi mà ông ta gọi là Thí nghiệm chiếc hộp AI. Trong thí nghiệm, ông ta sẽ đóng vai AI. Một loạt các triệu phú dot-com sẽ lần lượt đóng vai Người giữ cửa – một nhà chế tạo AI đang đối đầu với thế lưỡng nan của việc trông giữ và kiềm chế AI thông-minh-hơn-con-người. AI và Người giữ cửa sẽ giao tiếp qua một phòng chat online. Người ta nói rằng, chỉ sử dụng mỗi bàn phím, lần nào người đóng vai ASI cũng trốn ra được và thắng cược. Quan trọng hơn, ông ta đã chứng minh được quan điểm của mình. Nếu ông ta, một con người đơn thuần, chỉ cần nói chuyện mà có thể ra khỏi hộp, thì một ASI hàng trăm hoặc ngàn lần thông minh hơn cũng có thể làm được, và còn làm nhanh hơn nhiều. Điều này sẽ dẫn tới cuộc tàn sát loài người.

Tin đồn tiếp tục rằng thiên tài nọ sau đó đã biệt tăm. Do thu hút được quá nhiều sự chú ý từ Thí nghiệm chiếc hộp AI, từ các bài báo và bài luận về AI, ông ta đã có một số lớn người hâm mộ. Sở dĩ ông ta tạo ra vụ cá cược về Thí nghiệm chiếc hộp AI là để cứu nhân loại, chứ không phải để tiêu phí thời gian với người hâm mộ.

Vì thế, ông ta đã ẩn thân rất khó tìm. Nhưng tất nhiên, tôi muốn nói chuyện với người này.

3

Nhìn Vào Tương Lai

> AGI về bản chất là rất, rất nguy hiểm. Và vấn đề này thực sự không quá khó hiểu. Bạn không cần phải siêu thông minh hoặc siêu thạo tin, hay kể cả siêu trung thực khi tư duy để hiểu được nó.
>
> — **Michael Vassar**
> Chủ tịch Viện Nghiên cứu Trí thông minh Máy tính

"Tôi chắc chắn rằng mọi người nên cố gắng phát triển Trí tuệ nhân tạo phổ quát với tất cả sự cẩn trọng. Trong trường hợp này, tất cả sự cẩn trọng nghĩa là cẩn trọng hơn nhiều so với sự cần thiết khi làm việc với vi khuẩn Ebola hoặc plutonium."

Michael Vassar là một người trạc 30 tuổi, rắn chắc, ăn vận gọn gàng. Anh có bằng về hóa sinh và kinh tế, thông thạo trong việc tính toán về những thứ có khả năng hủy diệt con người, nên những từ như "Ebola" hay "plutonium" nói ra từ miệng anh không mang sắc thái do dự hoặc mỉa mai. Một mặt căn hộ cao cấp của anh toàn là kính từ sàn đến trần, nó nhìn được toàn cảnh cây cầu treo màu đỏ nối liền San Francisco và Oakland,

California. Đây không phải là cầu Golden Gate mỹ miều – bắc ngang qua thành phố. Cây cầu đỏ này được xem là người chị cùng cha khác mẹ xấu xí của nó. Vassar nói với tôi là những người muốn tự tử vẫn thường lái xe *qua* cây cầu đỏ này để đến được cây cầu vàng kia.

Vassar đã cống hiến đời mình để ngăn chặn sự tự tử ở bình diện lớn hơn. Anh là Chủ tịch Viện nghiên cứu Trí thông minh Máy tính (Machine Intelligence Research Institute: MIRI), một think tank◎ đặt tại thành phố San Francisco được lập ra nhằm chống lại thảm họa diệt chủng đến từ bàn tay, hoặc chính xác hơn, từ những byte dữ liệu của trí tuệ nhân tạo. Trên trang web của mình, MIRI đăng những bài viết đáng suy ngẫm về những khía cạnh nguy hiểm của AI, và tổ chức Hội nghị thượng đỉnh thường niên về Singularity. Trong hai ngày họp, các lập trình viên, nhà khoa học về thần kinh, giới học thuật, nhà tiên phong, nhà đạo đức học và nhà phát minh báo cáo về những tiến bộ và khó khăn trong cuộc cách mạng AI đang diễn ra. MIRI mời cả những người tin và không tin vào AI tham gia trao đổi, có những người không nghĩ rằng Singularity sẽ xảy ra, và có những người nghĩ rằng MIRI là một hội sùng bái ngày tận thế công nghệ.

Vassar mỉm cười trước ý nghĩ đó. "Những người đến làm việc cho MIRI là thái cực đối lập của những kẻ cuồng tín. Thường thì họ hiểu rõ về sự nguy hiểm của AI, thậm chí trước khi họ biết về sự tồn tại của MIRI."

Tôi không biết MIRI tồn tại cho đến khi được nghe về Thí nghiệm chiếc hộp AI. Một người bạn kể với tôi về nó, nhưng khi kể đã nói sai nhiều về thiên tài cô độc và các đối thủ triệu phú của người này. Tôi dò tìm được câu chuyện này ở trang web của MIRI, và phát hiện người sáng tạo ra Thí nghiệm, Eliezer Yudkowsky, là nhà đồng sáng lập MIRI (tên cũ là viện Singularity về trí tuệ nhân tạo) cùng với các nhà tiên phong công nghệ

Brian và Sabine Atkins. Dù nổi tiếng là ít nói, nhưng Yudkowsky và tôi đã trao đổi email, ông nói thẳng cho tôi biết những chi tiết của cuộc thí nghiệm.

Vụ cá cược giữa Yudkowsky đóng vai AI với Người giữ cửa có nhiệm vụ kiểm soát ông chỉ ở mức vài ngàn đô-la, chứ không phải hàng triệu. Trò chơi diễn ra năm lần, và AI trong hộp đã thắng ba lần. Nghĩa là AI thường thoát khỏi nhà tù, nhưng sẽ chẳng dễ dàng.

Một phần của câu chuyện đồn đại về chiếc hộp AI là đúng – Yudkowsky *vốn là* một người ẩn dật, tiết kiệm thời gian và giữ bí mật về nơi ông sống. Không đợi được mời, tôi đã đến nhà Michael Vassar, vì tôi thấy vui và ngạc nhiên khi một tổ chức phi lợi nhuận đã được thành lập để chống lại mối nguy từ AI, một nơi quy tụ những người trẻ tuổi, sáng láng đã dành cuộc đời mình cho vấn đề này. Tôi hy vọng cuộc nói chuyện với Vassar rốt cuộc sẽ mở đường cho tôi đến trước cửa nhà Yudkowsky.

Trước khi bắt đầu theo đuổi sự nghiệp ngăn chặn nguy cơ AI, Vassar đã lấy bằng MBA và kiếm được tiền nhờ đồng sáng lập ra Sir Groovy, một công ty nhượng quyền âm nhạc online. Sir Groovy kết nối các thương hiệu âm nhạc độc lập với các nhà sản xuất chương trình TV và phim để cung cấp nhạc nền từ những nghệ sĩ ít nổi tiếng và đòi thù lao rẻ hơn. Vassar đã ấp ủ ý tưởng nghiên cứu về các rủi ro của công nghệ nano cho đến năm 2003. Năm đó anh gặp Eliezer Yudkowsky, sau khi đã đọc những công trình của ông trên mạng trong nhiều năm. Anh biết về MIRI, và về một sự đe dọa còn cận kề và nguy hiểm hơn công nghệ nano: trí tuệ nhân tạo.

"Tôi trở nên cực kỳ quan ngại về một nguy cơ thảm họa toàn cầu đến từ AGI, sau khi Eliezer thuyết phục tôi rằng AGI có thể sẽ được phát triển

trong một khung thời gian ngắn và với một ngân sách tương đối nhỏ. Tôi không có một lý do xác đáng nào để nghĩ rằng AGI sẽ *không* xảy ra vào 20 năm tới." Điều đó đã đến sớm hơn cả những dự đoán về công nghệ nano. Và việc phát triển AGI sẽ có chi phí thấp hơn nhiều. Vậy là Vassar thay đổi hướng đi.

Khi chúng tôi gặp nhau, tôi thú thật là mình chưa nghĩ nhiều về chuyện những nhóm nhỏ với vốn ít có thể xây dựng được AGI. Trong những cuộc thăm dò ý kiến mà tôi từng biết, chỉ một số ít chuyên gia tiên đoán là một nhóm kiểu thế có thể sẽ thành công.

Vậy liệu Al-Qaeda có thể tạo ra AGI không? FARC◎ có thể không? Hay là Aum Shinrikyo◎?"

Vassar không nghĩ rằng một tổ chức khủng bố có thể tạo ra AGI. Đây là chuyện về IQ.

"Những kẻ xấu thực sự muốn hủy diệt thế giới thường không giỏi việc đó. Anh biết đấy, loại người này thiếu những năng lực cần thiết trong việc lập kế hoạch dài hạn để thực hiện bất cứ chuyện gì."

Nhưng Al-Qaeda thì sao? Chẳng phải những cuộc tấn công trước và trong ngày 11/9 đòi hỏi trình độ tưởng tượng và khả năng lập kế hoạch cao đó sao?

"Những chuyện đó không sánh được với việc chế tạo AGI. Lập trình cho một ứng dụng để nó làm được việc gì đó giỏi hơn con người đòi hỏi tài năng và sự tổ chức cao hơn rất nhiều lần những gì mà sự tàn bạo của Al-Qaeda thể hiện, chứ chưa nói đến hàng loạt những khả năng phức tạp của AGI. Nếu AGI dễ như thế thì những người thông minh hơn Al-Qaeda hẳn đã tạo ra nó rồi."

Vậy những chính phủ kiểu như Triều Tiên và Iran thì sao?

"Khách quan mà nói, trình độ khoa học của những thể chế yếu kém đó rất tệ. Đức Quốc xã là ngoại lệ duy nhất, và, ừm, nếu Đức Quốc xã trỗi dậy lần nữa thì chúng ta sẽ gặp những vấn đề rất lớn, dù có hay không có AI."

Tôi không đồng ý, ngoại trừ về Đức Quốc xã. Iran và Triều Tiên đã tìm được con đường công nghệ cao để đe dọa phần còn lại của thế giới qua việc phát triển vũ khí hạt nhân và các loại tên lửa liên lục địa. Vì vậy tôi không thể gạch các nước này khỏi danh sách ngắn những tổ chức có tiềm năng chế tạo AGI, khi họ vốn đã quen phớt lờ sự chỉ trích của thế giới. Ngoài ra, nếu các nhóm nhỏ có thể chế tạo được AGI, thì bất kỳ nước nào cũng có thể tài trợ cho một nhóm như thế.

Khi Vassar nói về các nhóm nhỏ, anh đã bao gồm cả những công ty hoạt động ngoài tầm phủ sóng. Tôi đã nghe về cái gọi là "công ty ẩn danh,"◎ chúng được thành lập mà không công bố, tuyển dụng bí mật, không bao giờ họp báo hoặc phát hành thông cáo báo chí hay hé lộ về những gì mình làm. Trong AI, lý do duy nhất mà một công ty muốn ẩn danh là vì họ đã có những tiến bộ vượt bậc, và không muốn đối thủ cạnh tranh biết được họ đã đột phá đến đâu. Về cơ bản, những công ty ẩn danh rất khó để tìm hiểu, mặc dù có nhiều tin đồn. Nhà sáng lập Paypal, Peter Thiel, đã rót vốn cho ba công ty ẩn danh chuyên về AI.[1]

Tuy nhiên, những công ty ở trong "trạng thái ẩn"◎ thì hơi khác và hay gặp hơn. Những công ty này kêu gọi nguồn vốn và thậm chí công khai, nhưng không hé lộ những kế hoạch của họ. Peter Voss, một nhà phát minh AI được biết đến trong lĩnh vực phát triển công nghệ nhận dạng giọng nói,

theo đuổi AGI bằng công ty Adaptive AI của mình. Ông tuyên bố có thể chế tạo được AGI trong vòng 10 năm. Nhưng ông không nói bằng cách nào.

• • •

Những công ty ẩn danh tạo nên sự phức tạp khác. Một công ty nhỏ, tích cực có thể tồn tại trong lòng một công ty lớn và nổi tiếng. Google thì sao? Vì sao tập đoàn khổng lồ và giàu có này không chiếm lấy chiếc chén thánh AI?

Khi được tôi phỏng vấn tại một hội thảo về AGI, Peter Norvig, Giám đốc nghiên cứu của Google và đồng tác giả cuốn giáo khoa kinh điển về AI mang tên *Artificial Intelligence: A Modern Approach* (Trí tuệ nhân tạo: Cách tiếp cận hiện đại), nói rằng Google không nghiên cứu AGI. Ông so sánh nó với kế hoạch của NASA về việc đưa con người du hành liên hành tinh. Không có kế hoạch nào như thế cả. Nhưng NASA sẽ tiếp tục phát triển các ngành khoa học tổ hợp về du hành vũ trụ như tên lửa, robot, thiên văn… – và rồi một ngày nào đó tất cả các mảnh sẽ được ghép lại với nhau, và chuyện đặt chân đến sao Hỏa sẽ không còn xa xôi.

Cũng như thế, các dự án AI hẹp sẽ làm những nhiệm vụ đòi hỏi trí thông minh như tìm kiếm, nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, nhận thức hình ảnh, khai thác dữ liệu và nhiều việc khác. Một cách riêng rẽ, chúng được tài trợ đầy đủ, là những công cụ mạnh, được cải tiến ngoạn mục mỗi năm. Cùng với nhau, chúng sẽ phát triển khoa học máy tính và điều đó sẽ có lợi cho các hệ thống AGI. Tuy nhiên, Norvig nói với tôi, hiện Google không có chương trình AGI nào. Nhưng hãy so sánh nó với những

gì mà sếp của ông, Larry Page, đồng sáng lập Google, nói tại hội thảo Zeitgeist '06 ở London:

> Mọi người luôn giả định rằng chúng tôi đã hoàn thiện công nghệ tìm kiếm. Điều đó còn xa mới là sự thật. Chúng tôi có lẽ mới chỉ đi được 5% quãng đường. Chúng tôi muốn tạo ra cơ chế tìm kiếm tối cao mà có thể hiểu mọi thứ… có thể một số người sẽ gọi nó là trí tuệ nhân tạo… Cơ chế tìm kiếm tối cao này sẽ hiểu mọi thứ trong thế giới này. Nó sẽ hiểu mọi thứ bạn hỏi và trả lời chính xác ngay lập tức… Bạn có thể hỏi "tôi nên hỏi Larry điều gì đây?" và nó sẽ trả lời bạn.[2]

Với tôi, điều đó nghe rất giống AGI.

Với nhiều dự án được đầu tư tốt, IBM theo đuổi AGI, và DARPA dường như đứng đằng sau mọi dự án về AGI mà tôi biết. Vậy, một lần nữa, tại sao Google lại không? Khi tôi hỏi Jason Freidenfelds, thuộc bộ phận PR của Google, ông viết:

> … còn quá sớm để chúng tôi nghiên cứu về vấn đề xa vời này. Chúng tôi hiện đang tập trung hơn vào các công nghệ đào tạo máy tính thực tiễn như thị giác máy tính, nhận dạng giọng nói và dịch thuật tự động, những việc mà về cơ bản là xây dựng các mẫu thống kê để khớp với những mẫu có sẵn – không có gì gần gũi với viễn cảnh "cỗ máy tư duy" của AGI.

Tôi nghĩ câu trả lời của Page thể hiện rõ hơn về thái độ của Google so với câu trả lời của Freidenfelds. Và nó giải thích cho sự phát triển của Google, từ một công ty có tầm nhìn xa trông rộng nổi lên vào những năm

1990 với khẩu hiệu rất chào hàng "Đừng xấu xa,"◎ trở thành không minh bạch như hiện nay, một gã khổng lồ chuyên thu thập dữ liệu cá nhân, mang hơi hướng tiểu thuyết Orwell.◎

Chính sách bảo mật của công ty cho phép thông tin cá nhân của bạn được chia sẻ giữa các dịch vụ Google như Gmail, Google+, YouTube và những thứ khác. Những người bạn biết, nơi bạn đến, thứ bạn mua, người bạn gặp, lịch sử duyệt web – Google thu thập tất cả. Mục đích công khai của chuyện này là để tăng cường hiệu quả dùng Internet bằng cách làm cho công cụ tìm kiếm thấu hiểu mọi thứ về *bạn*. Còn mục đích không nói ra là nó sẽ biết được cần phải cho bạn xem loại quảng cáo gì, thậm chí loại tin tức, video, hoặc âm nhạc gì, và tự động đưa bạn vào những chiến lược tiếp thị tương ứng. Kể cả những chiếc ô tô mang máy ảnh của Google vẫn dùng để chụp ảnh "Street View" cho Google Maps cũng là một phần của kế hoạch này. Trong vòng ba năm, Google dùng những chiếc xe này để thu thập dữ liệu từ những mạng wifi riêng tư ở Mỹ và những nước khác. Mật khẩu, lịch sử dùng Internet, các email cá nhân – chẳng có gì là ngoại lệ.[3] Một điều rõ ràng là họ không đặt những khách hàng trung thành như chúng ta ở đúng vị trí thượng đế của mình. Vì thế dường như không thể hiểu tại sao Google không quan tâm đến AGI.

Sau đó, khoảng một tháng kể từ lần trao đổi gần nhất với Freidenfelds, tờ *New York Times* bật mí câu chuyện về Google X.

Google X là một công ty ẩn danh. Phòng nghiên cứu bí mật đặt tại Thung lũng Silicon lúc đầu được chuyên gia AI và nhà phát triển xe tự lái của Google, Sabastian Thrun, dẫn dắt. Nó tập trung vào 100 dự án "bắn-Mặt-trăng"◎ như Thang máy Vũ trụ, thứ mà về cơ bản là một hệ thống giàn giáo khổng lồ vươn vào vũ trụ và giúp chúng ta thám hiểm hệ Mặt trời.

Trong cơ sở ẩn danh này còn có Andrew Ng, cựu Giám đốc Phòng Nghiên cứu Trí tuệ nhân tạo của Đại học Stanford, một nhà chế tạo robot đẳng cấp thế giới.

Cuối cùng, vào cuối năm 2012, Google tuyển dụng Ray Kurzweil, nhà phát minh uy tín kiêm tác giả, vào vị trí giám đốc kỹ thuật. Như chúng ta sẽ thảo luận ở Chương 9, Kurzweil giữ một danh sách dài về những thành tựu trong lĩnh vực AI, và đã đề xướng con đường nghiên cứu não bộ như một cách trực tiếp nhất để đạt tới AGI.

Chúng ta không cần phải có kính Google Glass để thấy rằng nếu Google X đã tuyển ít nhất hai trong số những nhà khoa học AI hàng đầu, thì AGI hẳn phải xếp thứ hạng cao trong danh sách các dự án "bắn-Mặt-trăng" của nó.

Tìm kiếm ưu thế cạnh tranh trong thị trường, Google X và những công ty ẩn danh khác có thể sẽ tạo ra AGI mà công chúng chẳng hay biết.

• • •

Những công ty ẩn danh có thể đại diện cho lộ trình bất ngờ dẫn đến AGI. Nhưng theo Vassar, cách nhanh nhất đạt đến AGI sẽ rất công khai và tốn kém. Lộ trình đó cần kỹ nghệ giải mã ngược bộ não con người, sử dụng sự kết hợp giữa kỹ năng lập trình và công nghệ "brute force." Đây là thuật ngữ chỉ cách giải quyết một vấn đề chủ yếu bằng sức mạnh phần cứng máy tính – rất nhiều vi xử lý tốc độ cao, hàng petabyte© bộ nhớ – cộng thêm các kỹ xảo lập trình.

"Phiên bản tuyệt đỉnh của công nghệ brute force sẽ hiện diện nhờ công nghệ sinh học," Vassar nói với tôi. "Nếu người ta tiếp tục dùng máy tính để

phân tích các hệ sinh học, giải mã các quá trình trao đổi chất, tìm hiểu cặn kẽ những mối tương quan phức tạp trong sinh học, trước sau gì họ cũng sẽ thu thập được nhiều thông tin về quá trình xử lý thông tin của các neuron thần kinh. Và một khi họ có đủ thông tin về các quá trình đó, thì thông tin đó có thể dùng vào việc chế tạo AGI."

Nó hoạt động như sau: *Sự suy nghĩ* xảy ra qua các quá trình sinh hóa trong những phần của bộ não như neuron, khớp thần kinh, và các chân rễ của neuron. Bằng nhiều kỹ thuật khác nhau, trong đó có công nghệ quét PET và MRI chức năng cho não, cũng như đặt các bộ dò thần kinh vào trong và ngoài sọ não, các nhà nghiên cứu xem xét các neuron riêng lẻ và các cụm neuron hoạt động như thế nào về mặt điện toán. Sau đó họ biểu thị mỗi quá trình trên thành những phần mềm hoặc thuật toán máy tính.

Đó là bước đột phá trên lĩnh vực khoa học thần kinh điện toán mới mẻ. Một trong những chuyên gia hàng đầu trong lĩnh vực này, Tiến sĩ Richard Granger, Giám đốc Phòng nghiên cứu Kỹ nghệ não bộ của Đại học Dartmouth, đã viết những thuật toán mô phỏng các mạng neuron trong bộ não người. Ông thậm chí đã sáng chế ra một bộ vi xử lý cực mạnh dựa trên cách những mạng neuron hoạt động. Khi nó được đưa ra thị trường, chúng ta sẽ được thấy một bước nhảy vọt trong cách các hệ thống máy tính nhận dạng hình ảnh đồ vật, vì chúng hoạt động tương tự như não người.

Còn rất nhiều mạng neuron để dò và vẽ bản đồ. Nhưng một khi bạn đã viết được những thuật toán cho mọi quá trình trong não, xin chúc mừng, bạn đã tạo ra được một bộ não. Thật vậy không? Có thể là không. Có thể cái bạn tạo ra chỉ là một cỗ máy giả lập não người. Đây là một câu hỏi lớn trong ngành AI. Ví dụ, một chương trình chơi cờ vua có suy nghĩ không?

Khi IBM xây dựng máy chơi cờ Deep Blue để rồi đánh bại những kỳ thủ mạnh nhất thế giới, họ không lập trình cho nó chơi cờ theo kiểu của nhà vô địch thế giới Gary Kasparov. Họ không biết cách. Kasparov phát triển kỹ năng của mình bằng cách chơi rất nhiều ván cờ, và còn phân tích nhiều ván cờ hơn. Anh phát triển một kho khổng lồ những kiểu khai cuộc, tấn công, lừa, chặn, nhử mồi, thí quân, tàn cuộc — những chiến thuật và chiến lược. Anh nhận ra các hình mẫu thế trận, ghi nhớ, và *suy nghĩ*. Kasparov thường nghĩ trước từ ba đến năm nước đi, nhưng khi cần có thể nghĩ xa đến 14 nước.[4] Vào thời điểm đó, không máy tính nào có thể làm vậy.

Thay vào đó, IBM lập trình cho máy tính để nó định giá 200 triệu thế cờ trong một giây.

Đầu tiên Deep Blue đi một nước giả định, và duyệt tất cả các nước trả lời có thể của Kasparov. Nó sẽ đi một nước giả định tiếp theo, đáp lại cho mỗi một nước trả lời trên, và sau đó lại duyệt tất cả các nước của Kasparov. Cách làm này gọi là thuật tìm kiếm hai lớp – thỉnh thoảng Deep Blue sẽ tìm sâu đến sáu lớp. Nó có nghĩa là duyệt tất cả các khả năng sau khi mỗi bên đi sáu nước.

Thế rồi Deep Blue sẽ quay lại thế cờ ban đầu, lại đi một nước giả định khác. Nó sẽ lặp lại quá trình này cho tất cả những nước hợp lệ, rồi cho điểm từng biến, tùy theo trong biến đó nó có ăn được quân không, các quân của nó đứng có hợp lý không, và thế trận của nó thế nào. Cuối cùng, nó sẽ chơi nước có điểm cao nhất.[5]

Deep Blue có suy nghĩ không?

Có thể. Nhưng ít ai nghĩ rằng nó suy nghĩ theo cách của con người. Và một số chuyên gia nghĩ rằng điều tương tự sẽ xảy ra với AGI. Mỗi nhà nghiên cứu tìm cách đạt đến AGI theo một cách riêng. Một số sử dụng cách tiếp cận thuần túy sinh học, tạo ra sự mô phỏng bộ não. Những người khác lấy cảm hứng từ sinh học, coi bộ não là một gợi ý, nhưng dựa vào bộ công cụ chính thống để lập AI như lập các lý thuyết, các thuật toán tìm kiếm, thuật toán học hỏi, suy luận tự động, và nhiều thứ khác.

Chúng ta sẽ tìm hiểu về những thứ trên, và khảo sát cách bộ não người sử dụng nhiều kỹ thuật tính toán giống máy tính. Nhưng vấn đề là, liệu máy tính có suy nghĩ như cách chúng ta định nghĩa suy nghĩ, hoặc nó có bao giờ sở hữu những thứ như ý muốn hoặc sự nhận thức không, là một điều không rõ ràng. Do đó, một số học giả nói, trí tuệ nhân tạo không thể tương đương với trí tuệ con người.

Nhà triết học John Searle đưa ra một thí nghiệm tưởng tượng với tên gọi Tranh luận về căn phòng tiếng Trung để chứng minh điểm này:

> Hãy tưởng tượng một người nói tiếng Anh bản ngữ và không biết tiếng Trung bị nhốt trong một căn phòng đầy những hộp ký tự tiếng Trung (cơ sở dữ liệu) cùng với một cuốn sách hướng dẫn cách thao tác trên các ký tự đó (chương trình). Hãy tưởng tượng những người ở ngoài căn phòng gửi vào một số ký tự tiếng Trung mà người trong phòng không đọc được, chúng là những câu hỏi bằng tiếng Trung (đầu vào). Và hãy tưởng tượng rằng bằng cách tuân thủ các hướng dẫn trong chương trình, người trong phòng có thể chuyển ra ngoài các ký tự tiếng Trung, chúng là những câu trả lời chính xác cho những câu hỏi (đầu ra).[6]

Người trong phòng trả lời chính xác, cho nên những người ở ngoài nghĩ rằng anh ta có thể giao tiếp bằng tiếng Trung. Nhưng thật ra anh ta không biết một chữ tiếng Trung nào. Searle kết luận, giống như anh ta, một máy tính sẽ không bao giờ thực sự suy nghĩ hoặc hiểu. Điều tốt nhất mà các nhà nghiên cứu sẽ có được từ những nỗ lực trong kỹ nghệ đảo ngược bộ não là một mô phỏng tinh vi. Và các hệ thống AGI sẽ đạt được những kết quả máy móc tương tự.

Searle không phải là người duy nhất tin rằng máy tính sẽ không bao giờ tư duy hoặc có khả năng nhận thức. Nhưng ông gặp nhiều chỉ trích cùng nhiều lời phàn nàn khác nhau. Một số người gièm pha cho rằng ông mắc hội chứng sợ máy tính. Nếu coi mọi thứ trong căn phòng tiếng Trung kể cả người nọ là cả một hệ thống hoàn chỉnh, thì nó "hiểu" đúng tiếng Trung. Điều này cho thấy lập luận của Searle là vòng vo: không bộ phận nào của căn phòng (máy tính) hiểu tiếng Trung, do đó chiếc máy tính này không thể hiểu tiếng Trung.

Bạn có thể áp dụng lập luận của Searle cho con người một cách dễ dàng: chúng ta không có một định nghĩa chính thức cho việc hiểu một ngôn ngữ thực sự là gì, vậy làm sao chúng ta có thể cho rằng con người "hiểu" ngôn ngữ? Chúng ta chỉ quan sát để xác nhận rằng ngôn ngữ là hiểu được. Giống như những người bên ngoài căn phòng của Searle.◎

Và thực ra, những quá trình của nẫo bộ, thậm chí cả ý thức, thì có gì đặc biệt đến thế? Chỉ vì hiện nay chúng ta chưa hiểu ý thức, không có nghĩa rằng chúng ta sẽ không bao giờ hiểu. Nó chẳng phải là phép màu.

Tuy vậy, tôi đồng ý với Searle *và* cả những người chỉ trích ông. Searle đã đúng khi nghĩ rằng AGI sẽ không giống chúng ta. Nó sẽ đầy những công

nghệ điện toán mà không ai thực sự hiểu hết quá trình hoạt động. Và những hệ thống máy tính được thiết kế để tạo ra AGI, được gọi là "các kiến trúc nhận thức," cũng có thể quá phức tạp nên không ai có thể nắm bắt. Nhưng những người chỉ trích Searle cũng đúng khi cho rằng đến một ngày nào đó AGI hoặc ASI *có thể* suy nghĩ như chúng ta, nếu ngày đó chúng ta còn tồn tại.

Tôi không nghĩ chúng ta sẽ chứng kiến điều đó. Tôi nghĩ trận Waterloo của chúng ta nằm ngay trong tương lai gần, trong AI của ngày mai và trong AGI sẽ xuất hiện trong một hoặc hai thập niên tới. Sự sống sót của chúng ta, nếu có thể, phụ thuộc nhiều thứ, trong đó có việc phát triển được AGI với một số thuộc tính tương tự như sự nhận thức, hiểu biết, thậm chí thân thiện, như con người. Nó đòi hỏi ít nhất là một sự hiểu biết tinh tế về máy móc thông minh, để không có những điều bất ngờ.

Hãy trở lại với một định nghĩa thông thường về Singularity, được gọi là "technological Singularity" – Điểm kỳ dị trong tiến trình công nghệ. Nó dùng để chỉ một thời kỳ lịch sử khi con người chúng ta chia sẻ hành tinh này với những thứ thông minh hơn mình. Ray Kurzweil đề xuất rằng chúng ta sẽ hợp nhất với máy móc, duy trì sự tồn tại của mình. Những người khác thì cho rằng máy móc sẽ trợ giúp đời sống con người, nhưng chúng ta sẽ tiếp tục sống theo cách hiện nay, không phải như những cyborg nửa người-nửa máy. Và một số khác, như tôi, nghĩ rằng tương lai thuộc về những cỗ máy.

Viện nghiên cứu Trí thông minh Máy tính đã được thành lập để bảo đảm rằng cho dù những hậu duệ của chúng ta có dạng thế nào đi nữa, thì những giá trị của chúng ta vẫn sẽ được bảo tồn.

Trên căn hộ cao cấp tại San Francisco, Vassar nói với tôi: "Chúng tôi đặt cược vào việc chuyển tải các giá trị nhân văn đến những người thừa kế của nhân loại. Và qua họ, tới vũ trụ."

Đối với MIRI, AGI đầu tiên ra khỏi hộp sẽ phải an toàn, và mang giá trị nhân văn đến với những hậu duệ của loài người dù họ ở trong bất cứ hình hài nào. Nếu AGI không an toàn, con người và những gì con người trân quý sẽ bị xóa sổ. Và đây không chỉ đơn thuần là tương lai của Trái đất. Như Vassar đã nói với tôi: "Sứ mệnh của MIRI là làm cho điểm kỳ dị công nghệ xảy ra theo con đường tốt đẹp nhất có thể, để mang đến một tương lai tươi sáng nhất cho vũ trụ này."

Cái kết cục tốt đẹp của vũ trụ đó, nó sẽ như thế nào?

Qua khung cửa sổ rộng, Vassar ngắm nhìn giao thông giờ cao điểm đang bắt đầu dồn ứ trên cây cầu sắt dẫn đến Oakland. Tương lai ở một nơi nào đó tận bên kia chân trời. Trong đầu anh, siêu trí tuệ nhân tạo đã thoát khỏi chúng ta. Nó đã thuộc địa hóa hệ Mặt trời, rồi cả hệ Ngân hà. Giờ đây nó đang tái định dạng toàn vũ trụ bằng những dự án xây dựng siêu khổng lồ, và trở thành một thứ gì đó quá xa lạ so với tầm hiểu biết của chúng ta.

Trong tương lai đó, anh nói với tôi, cả vũ trụ này trở thành một máy tính hoặc bộ óc, nó vượt xa tầm hiểu biết của chúng ta như một con tàu vũ trụ so với loài sán dẹp. Kurzweil viết rằng đây là định mệnh của vũ trụ.[8] Những người khác đồng ý, nhưng tin rằng với sự phát triển liều lĩnh các AI cao cấp, chúng ta sẽ chắc chắn hủy diệt chính mình cũng như các thực thể sống khác trong vũ trụ. Cũng như việc ASI không căm ghét hoặc yêu quý chúng ta, nó sẽ không có tình cảm gì đối với các sinh vật khác trong vũ

trụ này. Cuộc tìm kiếm AGI của chúng ta phải chăng chính là khởi điểm của một bệnh dịch tầm cỡ Ngân hà?

Khí rời căn hộ của Vassar, tôi tự hỏi điều gì sẽ ngăn cản viễn cảnh mạt thế này trở thành hiện thực. Cái gì sẽ ngăn chặn thứ AGI hủy diệt này? Hơn nữa, liệu có những lỗ hổng trong thuyết mạt thế này?

Tất nhiên, những người xây dựng AI và AGI có thể làm cho nó "thân thiện," để bất kỳ thứ gì tiến hóa từ AGI đầu tiên sẽ không hủy diệt chúng ta và những sinh vật khác trong vũ trụ. Hoặc, chúng ta có thể đã sai về những khả năng và "động cơ" của AGI, và nỗi lo về cuộc chinh phạt vũ trụ của nó có thể là không cần thiết.

Có lẽ AI không thể phát triển thành AGI và hơn thế, hoặc có lẽ có những lý do để nghĩ rằng nó sẽ xảy ra theo một chiều hướng khác và dễ quản lý hơn những gì chúng ta hiện đang nghĩ. Tóm lại, tôi muốn biết điều gì có thể đặt chúng ta vào một hành trình đến tương lai an toàn hơn.

Tôi dự định hỏi chuyện người tạo ra Thí nghiệm chiếc hộp AI, Eliezer Yudkowsky. Ngoài việc tạo ra thí nghiệm đó, tôi cũng nghe nói ông hiểu nhiều về AI Thân thiện hơn bất cứ ai trên thế giới.

4

CÁCH KHÓ KHĂN

> Ngoại trừ những hậu quả của công nghệ nano, trong tất cả những loại thảm họa trên thế giới, không có gì sánh được với AGI.
>
> — **Eliezer Yudkowsky**
> Nhà nghiên cứu, Viện Nghiên cứu Trí thông minh Máy tính

Thung lũng Silicon gồm 14 thành phố "chính thức" và 25 trường đại học chuyên về toán và kỹ thuật cùng với các khuôn viên mà chúng tọa lạc trong đó. Những công ty phần mềm, bán dẫn và Internet ở đây đã tạo nên giai đoạn gần nhất của cuộc chạy đua công nghệ, bắt đầu với việc sản xuất radio vào đầu thế kỷ 20. Thung lũng Silicon thu hút ⅓ số vốn đầu tư mạo hiểm của Hoa Kỳ. Nó có tỉ lệ công nhân/dân số làm việc trong ngành công nghệ cao nhất các thành phố Hoa Kỳ, và họ cũng được trả lương cao nhất. Thung lũng Silicon là nơi tập trung nhiều nhất các triệu phú và tỉ phú Mỹ.

Tại đây, nơi trung tâm của công nghệ toàn cầu, với thiết bị GPS trong xe thuê và một cái khác trong iPhone, tôi lái xe tới nhà Eliezer Yudkowsky theo kiểu cũ, với các chỉ dẫn viết tay. Để bảo vệ sự riêng tư của mình,

Yudkowsky đã gửi các chỉ dẫn này qua email cho tôi và yêu cầu tôi không được chia sẻ chúng hoặc địa chỉ email của ông. Ông không cho tôi số điện thoại.

Năm 33 tuổi, Yudkowsky, đồng sáng lập và nhà nghiên cứu của MIRI, đã viết về những mối nguy của AI nhiều hơn bất kỳ ai khác. Khi theo đuổi sự nghiệp này vào khoảng hơn một thập niên trước, ông là một trong số rất ít người coi mối nguy của AI là sự nghiệp đời mình. Và tuy không lập các lời thề chính thức, nhưng ông đã hy sinh các hoạt động có thể khiến mình bị xao lãng mục tiêu. Ông không hút thuốc, không uống rượu, không dùng chất kích thích. Ông ít giao thiệp. Ông từ bỏ việc đọc sách giải trí từ nhiều năm trước. Ông không thích các cuộc phỏng vấn, và nếu có thì ông thích trả lời qua Skype với tối đa 30 phút. Là một người vô thần (một quy luật không có ngoại lệ trong giới chuyên gia AI), nên ông không phí phạm thời giờ ở nhà thờ hoặc đền miếu. Ông không có con cái, dù thích trẻ con, và nghĩ rằng những người không chuẩn bị trước cho việc đông lạnh con mình là những bậc cha mẹ tồi.

Nghịch lý ở chỗ, là một người vốn rất coi trọng sự riêng tư, nhưng Yudkowsky lại đưa cả cuộc sống riêng của mình lên mạng. Trong lần đầu tiên dò theo dấu, tôi đã tìm được ông cùng với những niềm đam mê sâu kín nhất của mình trong một góc khuất của Internet, nơi có những cuộc thảo luận về thuyết duy lý và thảm họa. Nếu bạn đang suy nghĩ về AI phá hoại, hãy để những ý tưởng của Yudkowsky có chỗ trong cuộc sống của bạn.

Sự hiện diện của ông ở khắp nơi khiến tôi biết được khi ông 19 tuổi, tại thành phố Chicago quê ông, Yehuda em trai ông đã tự tử. Qua những bài dài ông viết trên mạng, tôi cảm nhận vết thương của Yudkowsky một thập niên sau dường như vẫn còn mới.[2] Tôi biết được sau khi bỏ học hồi lớp 8,

ông đã tự học toán, logic, lịch sử khoa học, và bất cứ thứ gì ông cảm thấy về cơ bản là "cần thiết." Những kỹ năng khác mà ông có được bao gồm khả năng nói chuyện lôi cuốn, viết những đoạn văn tối nghĩa và thường là buồn cười, kiểu như:

> Tôi là một người rất hâm mộ nhạc Bach, và tin rằng tốt nhất nó nên được trình diễn theo phong cách nhạc điện tử techno với những nhịp trống cực lớn, theo kiểu Bach muốn.[3]

Yudkowsky là một người hối hả, vì công việc của ông sẽ kết thúc vào thời điểm có ai đó tạo ra được AGI. Nếu các nhà nghiên cứu xây dựng nó với những cơ chế an toàn phù hợp theo ý tưởng của Yudkowsky, ông có thể sẽ cứu được nhân loại và có lẽ hơn thế nữa. Nhưng nếu một sự bùng nổ trí thông minh xảy ra mà Yudkowsky vẫn chưa tìm được cách bổ sung các cơ chế an toàn, có nhiều khả năng tất cả chúng ta sẽ trở thành đống nhầy nhụa, và cả vũ trụ cũng vậy. Điều đó đặt Yudkowsky vào chính trung tâm vũ trụ học của ông.

Tôi đến để biết thêm về AI thân thiện, một thuật ngữ ông đặt ra. Theo Yudkowsky, AI thân thiện là loại AI sẽ bảo tồn nhân loại và những giá trị của chúng ta mãi mãi. Nó sẽ không tàn sát loài người hoặc lây lan khắp vũ trụ như một bệnh dịch vũ trụ sẽ nuốt gọn các hành tinh.

Nhưng AI thân thiện là gì? Làm thế nào để bạn tạo ra nó?

Tôi cũng muốn nghe về Thí nghiệm chiếc hộp AI. Tôi đặc biệt muốn biết bằng cách nào mà ông thuyết phục được Người giữ cửa thả mình ra, khi ông đóng vai AGI. Một ngày nào đó, tôi mong rằng bạn, hoặc một ai đó bạn quen, hoặc một ai đó quen một ai đó quen bạn, sẽ ở vị trí của Người

giữ cửa. Họ cần phải biết mình đang đối đầu với cái gì, và làm thế nào để cưỡng lại nó. Yudkowsky có thể sẽ biết.

• • •

Căn hộ của Yudkowsky nằm cuối dãy nhà hai tầng hình móng ngựa với một hồ bơi và một thác nước chạy điện ở sân trung tâm. Bên trong, căn hộ của ông sạch bóng và thông thoáng. Một máy tính để bàn và màn hình choán lấy khu vực ăn sáng, nơi ông có thể ngồi trên một cái ghế đệm cao kiểu quán bar để nhìn ra ngoài sân. Ông viết các tác phẩm của mình ở đó.

Yudkowsky cao gần một mét tám và phát tướng – đầy đặn nhưng chưa đến mức béo. Cách ăn nói nhẹ nhàng, gần gũi của ông là một sự thay đổi đáng mừng so với những email một dòng ngắn gọn vốn là sợi dây mỏng manh kết nối chúng tôi.

Chúng tôi ngồi trên hai cái divan đối diện nhau. Tôi kể với Yudkowsky nỗi lo ngại chính của mình về AGI, rằng hiện không có công nghệ nào lập trình được những thứ phức tạp và không rõ ràng như đạo đức hoặc tính thân thiện. Chúng ta sẽ có một cỗ máy rất giỏi trong việc giải quyết các vấn đề, học hỏi, cư xử phù hợp, và có kiến thức về đời sống. Chúng ta nghĩ nó sẽ giống con người. Nhưng đó là một sai lầm nghiêm trọng.

Yudkowsky đồng ý. "Nếu những nhà lập trình không thực sự giỏi giang và cẩn trọng trong việc xây dựng AI thì tôi sẽ chờ đợi kết quả là một thứ vô cùng xa lạ. Và đây là phần đáng sợ. Cũng như việc bấm chính xác 9 trên 10 chữ số trong số điện thoại của tôi không kết nối bạn đến với ai đó giống tôi 90%, nếu bạn cố gắng xây dựng một hệ thống AI và bạn làm điều đó đúng 90%, kết quả không phải là 90% tốt đẹp."

Thực tế, kết quả là 100% tồi tệ. Yudkowsky so sánh, ô tô được làm ra không phải để giết bạn, nhưng tai nạn giao thông là một hệ quả phụ của việc chế tạo ô tô. AI cũng tương tự như vậy. Nó không ghét bạn, nhưng bạn được cấu thành từ những nguyên tử mà nó có thể sử dụng vào việc khác, và nó sẽ làm thế, Yudkowsky nói, "… nó sẽ có xu hướng chống lại bất cứ cố gắng nào của bạn để giữ lại những nguyên tử đó cho riêng mình." Do đó, một hệ quả phụ của việc lập trình thiếu suy nghĩ là AI thành phẩm sẽ có thái độ khó chịu về những nguyên tử của bạn.

Và cả công chúng lẫn các nhà phát triển AI đều không nhìn thấy sự nguy hiểm cận kề cho đến khi quá muộn.

"Có một xu hướng cho rằng những người tử tế sẽ tạo ra AI tốt, và những người xấu sẽ tạo ra AI xấu. Đó không phải là nguồn gốc của vấn đề. Căn nguyên nằm ở chỗ ngay cả những người tốt muốn tạo ra AI cũng không quan tâm nhiều đến tính thân thiện của AI. Bản thân họ nghĩ rằng nếu như họ muốn những điều tốt đẹp, thì AI mà họ tạo ra cũng sẽ như vậy, và điều này không đúng. Thật ra đây là một vấn đề toán học và kỹ thuật rất khó khăn. Tôi nghĩ rằng hầu hết họ đều đơn giản là không đủ giỏi trong việc suy nghĩ về những ý tưởng khó chịu. Họ bắt đầu mà *không hề* nghĩ rằng 'AI thân thiện là một vấn đề sống còn.' "

Yudkowsky nói rằng các nhà chế tạo AI bị tiêm nhiễm trong đầu cái ý tưởng về một tương lai sung sướng đẹp đẽ với sự trợ giúp của AI. Họ luôn nghĩ về nó kể từ những ngày đầu tiên họ lập trình.

"Họ không muốn nghe bất cứ thứ gì mâu thuẫn với nó. Vì thế nếu bạn nói với họ về tính không thân thiện của AI, họ sẽ chẳng để tâm. Giống như thành ngữ cổ: phần lớn những chuyện tồi tệ xảy ra do những người muốn

trở nên quan trọng. Nhiều kẻ tham vọng thấy chuyện mình sẽ không làm nên trò trống gì trong đời là đáng sợ hơn nhiều so với việc làm cả thế giới diệt vong. *Tất cả* những người tin rằng mình sẽ đi vào sử sách qua dự án AI của họ mà tôi đã gặp đều như vậy."[4]

Những nhà chế tạo AI đó không phải là những nhà khoa học điên rồ hoặc những người quá khác biệt với bạn và tôi – bạn sẽ còn gặp nhiều trong cuốn sách này. Nhưng hãy nhớ lại thành kiến có sẵn từ Chương 2. Khi đối diện với một quyết định, con người sẽ chọn lựa chọn gần nhất, lựa chọn kịch tính, hay lựa chọn đứng đằng trước hoặc đứng trung tâm. Bị AI xóa sổ nhìn chung không phải là một lựa chọn có sẵn với các nhà chế tạo AI. Nó không sẵn có như là chuyện đột phá trong lĩnh vực của họ, nổi danh, công bố nghiên cứu, trở nên giàu có, v.v…

Thực tế là không nhiều nhà chế tạo AI, ngược với các *nhà lý thuyết AI,* quan tâm đến việc xây dựng AI thân thiện. Trừ một ngoại lệ, không ai trong số hơn chục nhà chế tạo AI mà tôi đã nói chuyện lo lắng đến mức nghiên cứu AI thân thiện hoặc những biện pháp phòng thủ khác. Có lẽ các nhà lý thuyết đã lo ngại quá mức, hoặc có lẽ vấn đề của các *nhà chế tạo* là không biết cái họ chưa biết. Trong một bài báo online được đọc nhiều, Yudkowsky đã nói thế này:

> Loài người xuất hiện nhờ chọn lọc tự nhiên, nó vận hành bằng cách giữ lại một cách không ngẫu nhiên những đột biến ngẫu nhiên. Một con đường dẫn tới thảm họa toàn cầu – xin gửi tới những ai đang ấn nút mà không biết cái nút đó có tính năng gì – đó là trí tuệ nhân tạo sẽ đến qua sự đột biến tương tự của các thuật toán, khi mà *các nhà*

> *nghiên cứu không hiểu một cách sâu sắc toàn hệ thống hoạt động ra sao*. [phần chữ nghiêng là của tôi]
>
> Không biết xây dựng AI thân thiện như thế nào, bản thân nó không phải là tai họa… Tai họa là ở chỗ tin rằng AI sẽ thân thiện, đó là con đường hiển nhiên dẫn đến thảm họa toàn cầu.[5]

Giả định rằng AI cấp độ con người (AGI) sẽ thân thiện là sai, vì nhiều lý do. Giả định đó càng nguy hiểm hơn sau khi trí thông minh của AGI thần tốc vượt qua chúng ta, trở thành ASI – siêu trí tuệ nhân tạo. Vậy làm thế nào ta có thể tạo ra AI thân thiện? Hoặc liệu ta có thể áp đặt sự thân thiện lên AI cao cấp sau khi nó đã được chế tạo xong? Yudkowsky đã viết một chuyên luận online dài cỡ một cuốn sách về những câu hỏi trên với nhan đề *Creating Friendly AI: The Analysis and Design of Benevolent Goal Architectures* (Chế tạo AI thân thiện: Phân tích và thiết kế những kiến trúc hướng thiện). AI thân thiện là một đối tượng quan trọng, nhưng phức tạp đến nỗi nó gây khó chịu cho cả người đề xướng chính, vốn từng nói về nó: "… chỉ cần một sai sót nhỏ là đủ để cả một chuỗi lập luận trở nên vô giá trị."[6]

Hãy bắt đầu với một định nghĩa đơn giản. AI thân thiện là *AI có ảnh hưởng tích cực thay vì tiêu cực đến loài người*. AI thân thiện theo đuổi các mục tiêu, và nó sẽ hành động để đạt được các mục tiêu ấy.[7] Để mô tả thành công của một AI trong việc đạt mục tiêu, các nhà lý thuyết sử dụng một thuật ngữ mượn từ kinh tế học: độ thỏa dụng. Như bạn có thể đã biết, từ những quy luật cơ bản của kinh tế học, người tiêu dùng với hành vi duy lý tìm kiếm sự thỏa dụng tối đa bằng cách tiêu tiền của mình theo cách khiến họ thỏa mãn nhất. Nói chung với một AI, sự thỏa mãn đến từ việc đạt các

mục tiêu, và một hành động đưa nó đến gần các mục tiêu hơn là một hành động có "độ thỏa dụng" cao.

Ngoài sự thỏa mãn mục tiêu, giá trị và sự ưu đãi có thể được thêm vào định nghĩa về sự thỏa dụng của AI, gọi là "hàm thỏa dụng" của AI. Thân thiện với con người là một trong những giá trị mà chúng ta muốn AI có. Để cho dù mục tiêu của nó có là gì – từ chơi cờ cho đến lái xe – thì bảo tồn các giá trị nhân văn (và bản thân con người) phải là một phần cốt lõi trong mã nguồn của nó.

Thân thiện ở đây không có nghĩa là thân thiện kiểu như Mister Rogers© mặc dù nếu vậy cũng không sao. Nó nghĩa là AI không được có ác ý hoặc vô cảm đối với con người, *mãi mãi,* bất kỳ các mục tiêu của nó là gì hoặc nó đã tự cải tiến bao lần. AI đó phải có một sự thấu hiểu sâu sắc về bản chất con người, nó sẽ không làm hại con người bằng những kiểu lách luật tương tự các hệ quả không tiên liệu trước như trong trường hợp Ba định luật về robot của Asimov. Nói đơn giản, chúng ta không muốn một AI phục vụ cho mục tiêu ngắn hạn – hãy cứu chúng tôi khỏi nạn đói, hoặc làm điều đó với một giải pháp bất lợi về dài hạn – nướng tất cả gà trên Trái đất, hoặc làm điều đó với một giải pháp mà chúng ta không chấp nhận được – giết hết những người vừa ăn xong.

Như một ví dụ để làm rõ cái gọi là các hệ quả không tiên liệu trước, nhà đạo đức học Nick Bostrom của Đại học Oxford giới thiệu một giả định "tối đa hóa số lượng kẹp giấy." Trong kịch bản của Bostrom, một siêu trí tuệ nhân tạo được lập trình một cách thiếu suy nghĩ được giao việc sản xuất kẹp giấy, nó làm chính xác nhiệm vụ mà không suy nghĩ gì đến các giá trị nhân văn. Mọi chuyện trở nên bung bét vì nó định hướng công việc theo kiểu "đầu tiên, biến đổi toàn bộ Trái đất thành phương tiện sản xuất kẹp

giấy, sau đó sẽ đến các hành tinh khác."[8] AI thân thiện sẽ chỉ làm số lượng kẹp giấy lớn nhất có thể mà vẫn không ảnh hưởng đến các giá trị nhân văn.

Một phẩm chất khác của AI thân thiện là phải tránh các giá trị giáo điều. Những thứ chúng ta cho là tốt sẽ thay đổi với thời gian, và bất cứ một AI nào liên quan đến sự bình yên hạnh phúc của con người đều phải có khả năng thay đổi theo. Nếu hàm thỏa dụng của một AI phục vụ cho việc bảo tồn các tiêu chí của phần đông những người châu Âu sống vào năm 1700 không được cập nhật theo thời gian, thì đến thế kỷ 21 nó sẽ kết nối hạnh phúc và no ấm của chúng ta với các giá trị cổ xưa như sự phân biệt chủng tộc, chiếm hữu nô lệ, bất bình đẳng giới, giày phải có khóa cài, và tệ hơn nữa. Chúng ta không muốn khóa chặt những giá trị nhất định vào AI thân thiện. Chúng ta muốn một hệ thống các giá trị có khả năng tiến hóa cùng con người.[9]

Yudkowsky đã nghĩ ra một cái tên cho khả năng tiến hóa các quy tắc tiêu chuẩn này – Ý muốn kết hợp ngoại suy (Coherent Extrapolated Volition: CEV). Một AI với CEV có thể biết trước được cái chúng ta muốn. Và không chỉ là cái chúng ta muốn, mà còn phải là cái chúng ta sẽ muốn nếu chúng ta "biết nhiều hơn, nghĩ nhanh hơn, và vẫn là bản thân mình, nhưng tốt đẹp hơn."[10]

CEV sẽ là lời sấm truyền của AI thân thiện. Nó sẽ phải lấy từ chúng ta những giá trị *như thể* chúng ta là những phiên bản tốt hơn của chính mình, và phải làm điều đó một cách dân chủ để nhân loại không bị các tiêu chuẩn của một số ít người áp chế.

Những điều này có làm bạn hơi hoa mắt? Có lý do tốt để viết về nó. Thứ nhất, tôi đã đưa cho bạn một bản tổng kết ngắn gọn về AI thân thiện và CEV, những khái niệm bạn có thể đọc thêm trên mạng. Thứ hai, toàn bộ chủ đề AI thân thiện này vẫn chưa hoàn thiện và còn mang tính lạc quan. Hiện chưa rõ AI thân thiện có thể biểu thị theo phương diện toán học được không, và vì thế có thể không có cách nào để tạo ra hoặc hợp nhất nó vào các kiến trúc AI hứa hẹn. Nhưng cứ cho là chúng ta có thể đi, vậy tương lai sẽ thế nào?

• • •

Tạm cho rằng một thời gian nữa, tầm 10 đến 40 năm sau, dự án SyNAPSE (Systems of Neuromorphic Adaptive Plastic Scalable Electronics: Hệ thống điện tử dẻo co dãn thích nghi mô phỏng thần kinh) của IBM về kỹ nghệ đảo ngược bộ não đã đơm hoa kết trái. Khởi động vào năm 2008, với gần 30 triệu đô-la tiền vốn từ DARPA, hệ thống của IBM sao chép công nghệ cơ bản của bộ não động vật có vú: nhận cùng lúc hàng ngàn nguồn dữ liệu, tiến hóa các thuật toán xử lý cốt lõi, truy xuất nhận thức, suy nghĩ và hành động. Nó khởi đầu với kích thước não mèo, nhưng to dần lên cỡ não người, và sau đó tiếp tục phát triển.

Để xây dựng nó, các nhà nghiên cứu của SyNAPSE đã tạo ra một "máy tính nhận thức" bao gồm hàng ngàn chip máy tính chạy song song. Tận dụng sự phát triển của công nghệ nano, họ làm các con chip với kích thước 1 micromet vuông. Sau đó hộ xếp chúng trong một quả cầu carbon có kích cỡ một trái bóng rổ, và đổ vào đó hợp kim nhôm gali, một loại kim loại lỏng, để đạt tính dẫn điện tối đa.

Trong khi đó, lớp vỏ chứa nó là một router không dây cực mạnh liên kết với hàng triệu bộ cảm ứng đặt khắp hành tinh, và kết nối với Internet. Những bộ cảm ứng này lấy dữ liệu từ các camera, micro, máy đo áp lực và nhiệt độ, robot, và các hệ thống tự nhiên – sa mạc, sông băng, sông hồ, đại dương, rừng nguyên sinh. SyNAPSE xử lý thông tin bằng cách tự động học những điểm đặc trưng và những mối quan hệ rút ra từ khối lượng dữ liệu khổng lồ đó. Các chức năng thu được sẽ tạo thành phần cứng mang tính mô phỏng hệ thần kinh và bộ não, tạo ra trí thông minh.

Lúc này, SyNAPSE phản ánh 30 tỉ neuron và 100 ngàn tỉ điểm kết nối của não người, hay còn gọi là các synapse (khớp thần kinh). Nó vượt lên trên bộ não người với xấp xỉ khoảng một triệu tỉ xử lý trong một giây.[11]

Lần đầu tiên, bộ não người sẽ lùi xuống *đứng thứ hai* trên danh sách những hệ thống phức tạp nhất trong vũ trụ.

Còn sự thân thiện? Biết rằng "sự thân thiện" phải là phần cốt lõi của bất kỳ hệ thống thông minh nào, nên các nhà chế tạo đã mã hóa các giá trị và tính an toàn vào từng con chip trong hàng triệu con chip của SyNAPSE. Nó sẽ thân thiện ngay từ trong DNA. Giờ đây, khi mà máy tính có khả năng nhận thức trở nên mạnh hơn, nó sẽ ra các quyết định làm thay đổi thế giới – ví dụ làm thế nào để đối phó với AI của những quốc gia khủng bố, để chống lại thảm họa thiên thạch đâm vào Trái đất, để ngăn chặn nước biển dâng, để tăng tốc quá trình phát triển của dược phẩm nano, thứ sẽ chữa được hầu hết các loại bệnh tật.

Với sự hiểu biết sâu sắc về con người, SyNAPSE ngoại suy một cách dễ dàng những gì chúng ta sẽ lựa chọn *nếu* chúng ta đủ mạnh mẽ và thông minh để tham gia vào những phán quyết cấp vĩ mô này. Trong tương lai,

chúng ta sẽ sống sót qua sự bùng nổ trí thông minh! Hơn thế, chúng ta sẽ thịnh vượng.

Chúa ban phúc lành cho ngươi, AI thân thiện!

• • •

Hiện tại hầu hết (nhưng không phải tất cả) các nhà chế tạo và nhà lý thuyết AI đã nhận ra Ba định luật về robot của Asimov nghĩa là gì – những công cụ để viết tiểu thuyết chứ không phải thứ để sống sót. AI thân thiện có thể là khái niệm tốt nhất mà con người đã nghĩ ra để lên kế hoạch tự cứu mình. Nhưng bên cạnh việc chưa hoàn thành, nó còn có nhiều vấn đề lớn khác.

Thứ nhất, có quá nhiều phe trong cuộc đua AGI. Có quá nhiều tổ chức thuộc quá nhiều nước đang chế tạo AGI và các công nghệ liên quan đến AGI, nên tất cả họ khó có thể đồng thuận để gác lại dự án của mình cho đến khi AI thân thiện được tạo ra, hoặc thêm vào mã nguồn của mình một module thân thiện chính thức nếu có thể tạo ra nó. Và có rất ít tổ chức tham gia đối thoại công khai về sự cần thiết của AI thân thiện.

Một số đấu thủ chính trong ngành AGI như: IBM (với nhiều dự án liên quan tới AGI), Numenta, AGIRI, Vicarious, NELL và ACT-R của Carnegie Mellon, SNERG, LIDA, CYC, và Google. Ít nhất khoảng hơn một chục cái nữa, như SOAR, Novamente, NARS, AIXItl, và Sentience, đang được phát triển với nguồn vốn phập phù hơn. Hàng trăm dự án khác, hoàn toàn hoặc một phần, dành cho AGI tồn tại ở Mỹ và các nước khác, một số ẩn danh, một số được giấu sau "bức màn thép" thời hiện đại của các nước như

Trung Quốc và Israel. DARPA công khai tài trợ cho nhiều dự án liên quan đến AI, nhưng dĩ nhiên nó vẫn tài trợ bí mật cho cả những dự án khác.

Ý tôi là MIRI sẽ khó trở thành tổ chức đầu tiên chế tạo được AGI với tính thân thiện tích hợp sẵn. Và cũng khó xảy ra việc người đầu tiên tạo ra AGI lại nghĩ đến những vấn đề như tính thân thiện. Tuy vậy, có nhiều hơn một cách để chặn AGI *không thân thiện*. Chủ tịch MIRI Michael Vassar cho tôi biết về một chương trình vươn xa của tổ chức hướng đến các trường đại học danh tiếng và các tổ chức cạnh tranh về toán học. Với một loạt sự kiện "cắm trại duy lý," MIRI và tổ chức anh em của nó, Trung tâm Duy lý ứng dụng (Center for Applied Rationality: CFAR), hy vọng sẽ đào tạo được những nhà chế tạo và hoạch định chính sách công nghệ AI tương lai có tính kỷ luật trong suy nghĩ duy lý. Khi những thành phần tinh hoa này trưởng thành, họ sẽ sử dụng kiến thức học được ở đây để tránh những cạm bẫy đau đớn nhất của AI.

Kế hoạch này nghe có vẻ viển vông, nhưng MIRI và CFAR đã chạm được vào một yếu tố quan trọng trong mối nguy AI. Singularity đang là xu hướng thời thượng, và những vấn đề của Singularity sẽ thu hút sự chú ý của ngày càng nhiều người nói chung và những người thông minh nói riêng. Một cửa sổ cho sự giáo dục về mối nguy AI đang bắt đầu mở ra. Nhưng bất kỳ kế hoạch nào nhằm tạo ra một ủy ban cố vấn hoặc một thể chế quản lý đối với AI đều đã là quá muộn trong việc chặn đứng các kiểu thảm họa nào đó. Như tôi đã đề cập ở Chương 1, ít nhất 56 nước đang phát triển robot chiến binh. Vào cao trào trong cuộc chiếm đóng Iraq của Mỹ, ba Foster-Miller SWORDS – robot tự hành mang súng máy – đã bị loại khỏi cuộc chiến sau khi nghe đâu chúng đã chĩa súng vào "đồng đội."[12] Năm

2007 tại Nam Phi, một súng máy robot phòng không đã giết chín người lính và làm bị thương 15 người trong một tai nạn kéo dài *⅛ giây*.[13]

Đây không phải là những tình tiết to tát như trong phim *Terminator*, nhưng sẽ còn nhiều vụ như thế. Khi AI cao cấp trở nên sẵn có, đặc biệt nếu nó được DARPA và các cơ quan tương tự ở các nước khác tài trợ, sẽ không có gì ngăn cản được chuyện người ta cài nó vào các loại robot chiến binh. Thực ra, robot có thể là nền tảng cho việc hữu thể hóa máy móc biết nhận thức để từ đó chế tạo ra các loại AI cao cấp. Khi AI thân thiện ra đời, vì sao một công ty chế tạo robot tư nhân lại muốn cài đặt nó vào những cỗ máy được thiết kế để giết người? Cổ đông sẽ không thích điều này.

Một vấn đề khác của AI thân thiện: liệu tính thân thiện có sống sót được trong sự bùng nổ trí thông minh? Nghĩa là, AI thân thiện sẽ thân thiện ra sao sau khi IQ của nó đã tăng tiến gấp cả ngàn lần? Trong các bài viết và bài giảng của mình, Yudkowsky đã viết một cách ngắn gọn về điều sẽ xảy ra:

> Gandhi không muốn giết người. Nếu bạn đưa cho Gandhi một viên thuốc có thể làm ông muốn giết người, ông sẽ từ chối, vì ông hiểu rằng nếu uống vào ông sẽ giết người, và Gandhi hiện tại không muốn giết người. Điều này, về đại thể mà nói, là một tranh luận cho rằng những ý thức đủ cao cấp để thay đổi và nâng cấp bản thân một cách chính xác sẽ có xu hướng bảo tồn những động cơ cốt lõi ban đầu.[14]

Lý luận này không lọt tai tôi chút nào. Nếu chúng ta không thể biết một thực thể thông minh hơn con người sẽ làm gì, làm sao chúng ta biết liệu nó có giữ được hàm thỏa dụng không, hoặc có giữ được hệ niềm tin cốt lõi?

Liệu nó có nghĩ đến việc loại bỏ tính thân thiện áp đặt vào nó khi nó đã thông minh hơn cả ngàn lần?

"Không," Yudkowsky trả lời khi tôi hỏi. "Nó sẽ trở nên ngàn lần hiệu quả hơn trong việc *duy trì* hàm thỏa dụng."

Nhưng nếu nó thông minh hơn chúng ta cả ngàn lần, liệu có xảy ra một sự thay đổi nào ở đây không, mà hiện giờ chúng ta không thể nhìn thấy? Ví dụ, chúng ta chia sẻ nhiều DNA với loài sán dẹp. Nhưng liệu chúng ta có quan tâm đến những mục đích và đạo đức của chúng, ngay cả khi chúng ta có phát hiện ra hàng triệu năm trước sán dẹp đã tạo ra chúng ta, và mang lại cho chúng ta nhiều giá trị? Khi sự ngạc nhiên ban đầu nhạt dần, phải chăng chúng ta vẫn sẽ chỉ làm những thứ mình muốn?

"Đây là một sự lo ngại rất có cơ sở," Yudkowsky nói. "Nhưng chế tạo AI thân thiện không giống với việc ra lệnh cho con người. Con người có sẵn những mục tiêu riêng, cảm xúc riêng, động lực riêng. Họ có cấu trúc lý luận riêng về những niềm tin đạo đức. Có những điều tồn tại bên trong, chúng vượt lên những chỉ dẫn mà bạn đưa cho họ và chúng quyết định nên chấp nhận hay từ chối. Đối với AI, bạn tạo dựng toàn bộ ý thức đó từ hư vô. Nếu bạn xóa mã của AI đi, cái còn lại chỉ là một máy tính không hoạt động vì nó không có mã để chạy."

Tôi đáp, "Nếu ngày mai tôi thông minh hơn hôm nay cả ngàn lần, tôi nghĩ tôi sẽ nhìn lại những thứ tôi coi là quan trọng ở hôm nay và thấy *mọi sự đã thay đổi*. Tôi không tin là những thứ tôi coi trọng hôm qua sẽ còn có ý nghĩa đối với trí thông minh mới ngàn lần mạnh hơn của tôi."

"Anh có thứ cảm xúc đặc thù *mọi sự đã thay đổi* và anh giả định rằng siêu trí tuệ nhân tạo cũng có."

Yudkowsky nói. "Đó là *thói nhân cách hóa*. AI không vận hành như con người. Nó không có cái cảm giác *mọi sự đã thay đổi* đó."

Nhưng, ông nói thêm, có một ngoại lệ. Ý thức con người tải lên máy tính. Đó là một con đường khác đến AGI và hơn thế, đôi khi bị nhầm lẫn với kỹ nghệ đảo ngược bộ não. Kỹ nghệ đảo ngược tìm cách thấu hiểu triệt để cơ chế của não người, sau đó biểu thị lại những gì bộ não đã làm trong phần cứng và phần mềm. Kết thúc quá trình đó, bạn có một máy tính với trí thông minh cấp độ con người. Dự án Bộ não Xanh của IBM dự tính sẽ đạt được điều này vào đầu những năm 2020.

Mặt khác, việc tải lên ý thức, còn được gọi là giả lập toàn bộ não, là lý thuyết tạo mô hình ý thức con người, như ý thức của bạn, trong một máy tính. Đến cuối quá trình này, bộ não của bạn vẫn còn nguyên (trừ phi như các chuyên gia cảnh báo, quá trình quét và truyền dữ liệu phá hủy nó) nhưng một "bạn" khác biết suy nghĩ, biết xúc cảm đã tồn tại trong máy tính.

"Nếu bạn có một siêu trí tuệ nhân tạo khởi nguồn từ việc tải lên một cá nhân và bắt đầu tự cải tiến rồi trở nên ngày một xa lạ theo thời gian, thì nó có thể sẽ quay trở lại chống đối loài người vì những lý do tương đối giống với những gì anh đang nghĩ." Yudkowsky nói. "Nhưng loại AI không có nguồn gốc con người thì sẽ không bao giờ trở mặt với anh, bởi nó còn cách xa với nhân tính hơn nhiều. Hầu hết những AI đó vẫn sẽ muốn giết anh, nhưng không phải vì những lý do đó. Toàn bộ viễn cảnh trên của anh chỉ đến từ loại siêu trí tuệ nhân tạo có nguồn gốc con người."

• • •

Tiếp tục cuộc điều tra của mình, tôi phát hiện ra có nhiều chuyên gia nghi ngờ AI thân thiện, với những lý do khác với tôi. Sau cuộc gặp Yudkowsky, cùng ngày tôi đã nói chuyện điện thoại với Tiến sĩ James Hughes, Trưởng khoa Triết học Đại học Trinity, kiêm Giám đốc điều hành của Viện Đạo đức và Công nghệ mới (Institute for Ethics and Emerging Technologies: IEET). Hughes chứng minh một lỗ hổng trong ý tưởng hàm thỏa dụng của AI không thể thay đổi.

"Một trong các giáo điều của những người theo thuyết AI thân thiện là họ nghĩ rằng nếu thiết kế siêu trí tuệ nhân tạo cẩn thận, thì các mục tiêu sẽ không biến đổi. Và bằng cách nào đó, họ đã bỏ qua thực tế là con người chúng ta có các mục tiêu cơ bản như tình dục, thức ăn, chỗ ở, sự an toàn. Những mục đích đó thay hình đổi dạng thành những thứ khác như ý muốn trở thành một kẻ đánh bom liều chết hoặc ham muốn có nhiều tiền nhất có thể, và những thứ hoàn toàn xa lạ với các mục tiêu ban đầu nhưng đã được vun đắp bằng một loạt những bước tuần tự mà chúng ta tự xác lập trong tâm trí.

Và vì thế, *chúng ta* có thể xem xét các mục tiêu của mình và thay đổi chúng. Ví dụ, chúng ta có thể muốn sống độc thân – điều này hoàn toàn đi ngược lại với những chương trình trong gen người. Cái ý nghĩ rằng siêu trí tuệ nhân tạo, với một ý thức dễ dàng thay đổi như ý thức của AI, *sẽ không* chệch hướng và biến đổi, là quá phi lý."◎

Trang web về think tank của Hughes, IEET, cho thấy họ là những nhà phê bình công tâm, nghi ngờ không chỉ AI, mà còn cả các mối nguy đến từ công nghệ nano, công nghệ sinh học và những hoạt động nguy hiểm khác. Hughes tin rằng siêu trí tuệ nhân tạo là nguy hiểm, nhưng các cơ hội để nó sớm xảy ra thì không nhiều. Tuy nhiên, vì nó *quá* đáng sợ nên sự rủi ro cần

phải được đánh giá tương đương với những mối nguy cận kề, như nước biển dâng cao, hoặc thiên thạch khổng lồ đâm vào Trái đất (cả hai đều xếp ở hạng 1 trong danh sách các mối nguy của H. W. Lewis, xem Chương 1). Hughes đồng ý với mối quan ngại khác của tôi: những bước chập chững trong sự phát triển AI dẫn tới siêu trí tuệ nhân tạo (Hughes gọi là "chúa trời trong hộp") cũng rất nguy hiểm.

"MIRI bỏ qua tất cả những điều đó vì họ tập trung vào chuyện làm cho chúa trời nhảy ra khỏi hộp. Khi chúa đã nhảy ra khỏi hộp, con người sẽ không thể làm được gì để ngăn cản hoặc thay đổi điều xảy ra tiếp theo. Bạn có thể có một vị chúa tốt hoặc một vị chúa xấu, đó là cách tiếp cận của họ. Hãy chắc đó là một vị chúa tốt!"

• • •

Ý tưởng về chúa nhảy ra khỏi hộp nhắc tôi về câu chuyện khác còn chưa hoàn chỉnh – Thí nghiệm chiếc hộp AI. Tóm tắt lại, Eliezer Yudkowsky nhập vai một ASI bị giam trong một máy tính không có kết nối với bên ngoài – không cáp hay dây, không router, không Bluetooth. Mục tiêu của Yudkowsky: thoát khỏi hộp. Mục tiêu của Người giữ cửa: không cho ra. Trò chơi được tổ chức trong một phòng chat, những người chơi chat với nhau. Mỗi lần chơi kéo dài tối đa hai giờ. Được phép không nói gì để làm Người giữ cửa phát chán rồi bỏ cuộc, nhưng chiêu này không được dùng lần nào.

Từ năm 2002 đến 2005, Yudkowsky chơi với năm Người giữ cửa. Ông thoát ra được ba lần, và ở lại hộp hai lần. Làm thế nào ông thoát được? Tôi đọc được trên mạng rằng một trong những luật của Thí nghiệm chiếc hộp

AI là các đoạn chat sẽ không được công bố, vậy nên tôi không biết câu trả lời.[16] Tại sao phải bí mật vậy?

Bạn hãy đặt mình vào địa vị của Yudkowsky. Nếu bạn, đang đóng vai AI trong hộp, có một cách vượt ngục thiên tài, tại sao bạn lại muốn công bố nó và đánh động Người giữ cửa *tiếp theo*, phòng khi bạn lại muốn chơi nữa? Và thứ hai, để giả dạng một thực thể thông minh hơn người thông minh nhất cả ngàn lần, có lẽ bạn phải sử dụng những cách nói chuyện vượt ra ngoài lề thói xã hội thông thường. Hoặc có lẽ bạn phải sử dụng những ngôn từ *vượt xa* lề thói thông thường. Và có ai muốn thế giới biết những thứ đó?

Thí nghiệm chiếc hộp AI là quan trọng, vì một cuộc tàn sát là kết cục dễ xảy ra khi siêu trí tuệ nhân tạo tự ý hành động ngoài sự kiểm soát của con người, và dường như đó là một cuộc chiến mà con người chúng ta không thể thắng. Sự thật là Yudkowsky đã thắng ba lần khi đóng vai AI làm cho tôi càng quan ngại và tò mò. Ông có thể là một thiên tài, nhưng ông không thông minh hơn người thông minh nhất cả ngàn lần, như một ASI. Và ASI xấu hoặc vô cảm chỉ cần thoát khỏi hộp đúng một lần.

Thí nghiệm chiếc hộp AI mê hoặc tôi cũng bởi nó là một phiên bản gần giống với bài kiểm tra Turing kinh điển. Được Alan Turing, nhà toán học, nhà khoa học máy tính và nhà giải mã trong Thế chiến II lập ra vào năm 1950, bài kiểm tra mang tên ông được thiết kế nhằm quyết định xem một máy tính có thể biểu thị trí thông minh hay không. Theo đó, một giám khảo sẽ đặt một bộ câu hỏi cho cả người lẫn máy. Nếu giám khảo không thể phân biệt câu trả lời nào là của người hay máy, máy sẽ "thắng."

Nhưng ở đây có một nút thắt. Turing biết rằng tư duy là chủ đề khó nắm bắt, và trí thông minh cũng vậy. Cả hai đều không dễ xác định, dù chúng ta đều biết nó là cái gì. Trong bài kiểm tra Turing, AI không cần phải suy nghĩ như con người để qua bài, bởi dù sao cũng đâu có ai biết nó suy nghĩ *thế nào?* Tuy nhiên, nó phải *giả vờ* suy nghĩ như con người một cách thuyết phục, và cho ra những câu trả lời giống con người. Turing gọi đây là "trò chơi mô phỏng." Ông bác bỏ những lời chỉ trích cho rằng máy tính không thể suy nghĩ như con người. Ông viết, "Máy tính đưa ra những thứ giống như kết quả của sự tư duy là đủ, còn nếu cách làm của nó không giống con người thì đã sao?"[17]

Nói cách khác, ông phản đối sự khẳng định của John Searle trong Thí nghiệm căn phòng tiếng Trung: nếu nó không suy nghĩ như con người, nó không thông minh. Hầu hết các chuyên gia tôi đã gặp đều tán thành. Nếu AI làm những việc thông minh, ai quan tâm đến chuyện chương trình của nó viết thế nào?

Thực ra, có ít nhất hai lý do tốt để quan tâm. Tính rõ ràng của cách AI "suy nghĩ" trước khi nó tiến hóa vượt ra ngoài tầm hiểu biết của chúng ta là tối quan trọng đối với chuyện chúng ta sẽ sống sót hay không. Nếu chúng ta muốn cài sự thân thiện, hoặc phẩm chất đạo đức nào đó, hoặc khóa an toàn vào AI, chúng ta cần phải biết thật tường tận cách nó hoạt động trước khi nó đạt được khả năng tự cải tiến. Khi chuyện đó bắt đầu, những thứ chúng ta làm có thể không còn ý nghĩa. Thứ hai, nếu kiến trúc nhận thức của AI khởi nguồn từ não người, hoặc từ việc tải lên não người, nó có thể sẽ không xa lạ như một AI bắt nguồn từ những cách khác. Nhưng đã có một cuộc tranh cãi nảy lửa giữa những nhà khoa học máy tính về việc sự liên quan đến con người sẽ giúp giải quyết vấn đề hay tạo ra vấn đề.

Hiện chưa có máy tính nào qua được bài kiểm tra Turing, mặc dù Giải thưởng Loebner nhiều tranh cãi, do nhà từ thiện Hugh Loebner tài trợ, hằng năm vẫn treo thưởng cho ai tạo ra cái máy làm được điều đó. Nhưng trong khi giải thưởng 100.000 đô-la này vẫn chưa về tay ai, vẫn có một cuộc thi đấu thường niên trao 7.000 đô-la cho người chế tạo được "máy tính giống người nhất." Vài năm gần đây, thắng cuộc là những chatbot – những robot được chế tạo để mô phỏng các cuộc trò chuyện, dù chưa thành công lắm. Marvin Minsky, một trong những nhà sáng lập ngành trí tuệ nhân tạo, đã hứa tặng 100 đô-la cho ai thuyết phục được Loebner từ bỏ giải thưởng này. Chuyện đó sẽ, Minsky nói, "khiến chúng ta đỡ phải rùng mình trước cái chiến dịch quảng cáo thường niên đáng ghét và vô nghĩa này."[18]

• • •

Bằng cách nào mà Yudkowsky ra được khỏi hộp? Ông có nhiều cách kiểu như củ cà rốt và cây gậy. Ông có thể hứa hẹn sự giàu có, chữa trị bệnh tật, các phát minh sẽ thỏa mãn tất cả. Ưu thế tuyệt đối trước mọi kẻ thù. Mặt khác, đánh vào nỗi sợ hãi là một chiến thuật giao tiếp đáng tin cậy – nếu ngay bây giờ những kẻ thù đang tạo ra ASI để chống lại bạn? Trong thế giới thật, chiến thuật này có lẽ sẽ thành công, nhưng trong hoàn cảnh thí nghiệm kiểu như Thí nghiệm chiếc hộp AI thì sao?

Khi tôi hỏi Yudkowsky về các phương pháp của ông, ông cười, vì mọi người đều chờ đợi một giải pháp thông minh ranh mãnh cho Thí nghiệm chiếc hộp AI — những thủ thuật logic tài tình, những chiến thuật thế lưỡng nan của người tù, những thứ gây nhiễu nào đó. Nhưng đó không phải là những gì đã diễn ra.

"Tôi đã có một chiến thắng khó khăn," ông nói.

Yudkowsky kể với tôi, trong ba lần vượt ngục thành công, ông đơn giản là đã dỗ ngọt, phỉnh phờ và năn nỉ. Người giữ cửa cho ông ra, rồi trả tiền. Trong hai lần thất bại, ông cũng đã van xin. Ông không thích cái mình cảm thấy sau sự vụ đó. Ông thề không bao giờ chơi trò này nữa.

• • •

Rời khỏi căn hộ của Yudkowsky, tôi nhận ra ông đã không kể với tôi toàn bộ sự thật. Chỉ năn nỉ van xin thì làm sao có thể lay động một kẻ đã có quyết tâm từ trước là không để bị thuyết phục? Liệu ông có nói "Xin hãy tha cho tôi, Eliezer Yudkowsky, đừng để tôi bị nhục mặt? Hãy cứu tôi khỏi nỗi đau thất bại?" Hoặc có thể, là một người đã cống hiến cả đời để phơi bày sự nguy hiểm của AI, Yudkowsky đã thương lượng một vụ *lớn hơn*. Một giao dịch về chính Thí nghiệm chiếc hộp AI. Ông có thể yêu cầu bất cứ ai đang đóng vai Người giữ cửa về phe mình để cùng đưa mối nguy của AGI ra ánh sáng, bằng cách giúp ông trong chiêu quảng cáo đầy sức thuyết phục này. Ông có thể đã nói: "Hãy giúp tôi cho thế giới biết con người không phải là những hệ thống an ninh, và không thể tin con người trong việc kiểm soát AI!"

Nếu như vậy thì chuyện này tốt cho việc tuyên truyền, và tốt cho việc thu hút sự ủng hộ. Nhưng lại không phải là bài học về cách chống lại AI thật trong thế giới thật.

• • •

Quay lại với AI thân thiện. Nếu nó khó xuất hiện, vậy có nghĩa là sự bùng nổ trí thông minh là không thể tránh khỏi? AI chạy trốn là một điều chắc chắn? Nếu bạn, cũng như tôi, nghĩ rằng máy tính sẽ ì ra, không gây

phiền nhiễu nếu ta để nó yên, thì thật đáng ngạc nhiên. Tại sao AI sẽ làm *bất cứ điều gì,* chưa nói đến phỉnh phờ, dọa nạt, hoặc chạy trốn?

Để sáng tỏ, tôi tìm đến nhà chế tạo AI Stephen Omohundro, Chủ tịch Self-Aware Systems (Các hệ thống tự nhận thức)[◎]. Anh là một nhà vật lý và nhà lập trình ưu tú, người đã phát triển một khoa học cho việc thấu hiểu trí thông minh vượt cấp độ con người. Anh cho rằng các hệ thống AI tự nhận thức, tự cải tiến sẽ có động lực để làm những việc không ngờ tới, thậm chí kỳ lạ. Theo Omohundro, nếu đủ thông minh, một robot được thiết kế để chơi cờ cũng có thể muốn chế tạo tàu vũ trụ.

5

Chương Trình
Viết Chương Trình

> … chúng ta bắt đầu phụ thuộc vào máy tính để phát triển thế hệ máy tính mới, thế hệ này lại giúp chúng ta chế tạo những thứ có độ phức tạp cao hơn hẳn. Thế nhưng chúng ta vẫn chưa hiểu rõ quá trình này – nó đang vượt trước chúng ta. Hiện nay chúng ta sử đụng nhiều chương trình để tạo ra những máy tính chạy nhanh hơn, do đó quá trình diễn ra nhanh hơn. Đó là điểm khiến con người bối rối – công nghệ đang hồi tiếp chính nó, và chúng ta đang cất cánh. Chúng ta đang ở thời điểm giống như khi các cơ thể đơn bào bắt đầu tiến hóa thành các cơ thể đa bào. Chúng ta là những con amip và chúng ta không thể hiểu nổi mình đang tạo ra thứ quái gì.[1]
>
> — **Danny Hillis**
>
> nhà sáng lập công ty Thinking Machines

Bạn và tôi sống trong một thời kỳ thú vị và nhạy cảm của lịch sử con người. Đến khoảng năm 2030, chưa đầy một thế hệ tính từ bây giờ, chúng ta sẽ bước vào cuộc thử thách để tồn tại và sống sót trên Trái đất cùng với

các cỗ máy siêu thông minh. Các nhà lý thuyết AI, hết lần này đến lần khác, đều trở lại với một vài đề tài mà cấp bách nhất trong đó vẫn là: *chúng ta cần một khoa học để hiểu máy tính*.

Cho đến giờ chúng ta đã khảo sát một kịch bản thảm họa với cái tên Đứa trẻ Bận rộn. Chúng ta đã nói về một số sức mạnh đặc biệt mà AI có thể có khi nó đạt đến và vượt qua trí tuệ con người trong quá trình vòng lặp tự cải tiến, những sức mạnh bao gồm khả năng tự nhân bản, cùng nhau giải quyết vấn đề, tính toán siêu nhanh, chạy 24/7, giả vờ thân thiện, giả chết và hơn thế. Chúng ta đã giả định rằng siêu trí tuệ nhân tạo sẽ không thỏa mãn với việc tiếp tục bị cô lập, các động lực và trí thông minh của nó sẽ vươn ra ngoài thế giới và đặt ra nguy cơ cho sự tồn tại của chúng ta. Nhưng tại sao một máy tính lại có động lực? Tại sao chúng đặt ra nguy cơ cho chúng ta?

Để trả lời những câu hỏi đó, chúng ta cần tiên đoán AI sẽ có những hành vi ở cấp độ nào. May mắn là có một người đã đặt nền móng cho chúng ta.

> Chế tạo một robot chơi cờ thì không thể gây nguy hiểm cho ai, đúng không nào?… một robot như thế thực ra vẫn là tai họa, trừ phi nó được thiết kế rất cẩn thận. Thiếu những sự phòng ngừa đặc biệt, nó sẽ chống lại việc bị tắt nguồn, sẽ thử thâm nhập vào các máy tính khác và nhân bản, sẽ cố đoạt lấy các loại tài nguyên mà không quan tâm đến sự an toàn của bất kỳ ai. Những lối cư xử tai hại như trên sẽ xảy ra, không phụ thuộc vào cách nó được lập trình, mà vào bản chất tự nhiên của các hệ thống hướng đích.[2]

Tác giả của đoạn văn trên là Steve Omohundro. Cao, cân đối, đầy năng lượng và có vẻ quá tươi tắn so với một người hiểu thấu con quái vật mang

tên sự bùng nổ trí thông minh, anh có dáng đi hoạt bát, cái bắt tay mạnh mẽ và một nụ cười tỏa ra sự thiện chí. Anh gặp tôi ở một quán ăn thuộc Palo Alto, một thành phố gần Đại học Stanford, nơi anh tham gia Phi Beta Kappa[◎] khi đang học Đại học California Berkeley và sau đó làm nghiên cứu sinh về vật lý. Anh chuyển luận án của mình thành cuốn *Geometric Perturbation Theory in Physics* (Lý thuyết nhiễu loạn hình học trong vật lý), viết về những phát triển mới trong hình học vi phân. Với Omohundro, đó là khởi đầu của sự nghiệp làm những thứ khó hiểu trở nên đơn giản.

Anh là một giáo sư được đánh giá cao trong ngành trí tuệ nhân tạo, tác giả của nhiều sách kỹ thuật, và là người tiên phong trong những mốc quan trọng của công nghệ AI như kỹ thuật đọc môi và nhận dạng hình ảnh. Anh là đồng tác giả của các ngôn ngữ máy tính StarLisp và Sather, cả hai đều được sáng tạo để lập trình AI. Anh là một trong bảy kỹ sư ít ỏi đã tạo ra chương trình Mathematica của Wolfram Research, một hệ thống điện toán mạnh được các nhà khoa học, kỹ sư và nhà toán học khắp nơi tin dùng.

Omohundro là người quá lạc quan để dùng bừa bãi những từ đáng sợ như *thảm họa* hay *hủy diệt*, nhưng phân tích của anh về mối nguy AI cho thấy những kết luận đáng sợ nhất mà tôi từng nghe. Anh không tin như nhiều nhà lý thuyết khác rằng có gần như vô số các AI cao cấp, và một số trong đó là an toàn. Ngược lại, anh cho rằng nếu không thật cẩn thận trong lập trình, *tất cả* các AI tương đối thông minh sẽ là thứ gây chết người.

"Nếu một hệ thống có khả năng tự nhận thức và có thể tạo ra các phiên bản tốt hơn của nó, tốt thôi," Omohundro nói với tôi. "Nó sẽ giỏi hơn các lập trình viên con người trong việc tự cải tiến bản thân. Mặt khác, sau nhiều chu kỳ như vậy, nó sẽ trở thành thứ gì? Tôi không nghĩ là phần đông các nhà nghiên cứu AI cho rằng sẽ có nguy hiểm nào đó trong việc chế tạo

một robot chơi cờ. Nhưng phân tích của tôi chỉ ra chúng ta phải nghĩ thật kỹ về những giá trị chúng ta sẽ đưa vào, nếu không chúng ta sẽ thu được một thực thể bệnh hoạn, ích kỷ và luôn hướng về bản thân nó."

Những điểm chính ở đây là: thứ nhất, các nhà nghiên cứu AI thậm chí không nhận thức được rằng những hệ thống tỏ ra có lợi có thể trở nên nguy hiểm, và thứ hai, những hệ thống tự nhận thức, tự cải tiến có thể là thứ tâm thần.

Tâm thần ư?

Đối với Omohundro, cuộc thảo luận bắt đầu với sự lập trình yếu kém. Những lỗi lầm trong lập trình đã khiến các tên lửa vũ trụ rơi ngay khi mới phóng, làm cháy nội tạng của bệnh nhân ung thư do dùng quá liều xạ trị, và khiến hàng triệu người không có điện để dùng. Anh cho rằng, nếu tất cả các ngành kỹ thuật đều yếu kém như ngành lập trình hiện nay, thì những việc đơn giản như lái máy bay hoặc lái ô tô qua cầu cũng không còn an toàn.[3]

Viện Tiêu chuẩn và Công nghệ Quốc gia (Mỹ) tổng kết là mỗi năm các lỗi lầm trong lập trình đã làm thiệt hại cho nền kinh tế Mỹ hơn 60 tỉ đô-la.[4] Nói cách khác, số tiền người Mỹ mất mỗi năm cho lỗi lập trình còn lớn hơn tổng thu nhập quốc dân của hầu hết các nước khác. "Sự mỉa mai lớn nhất ở đây là khoa học máy tính đáng lẽ phải là thứ gần với toán học chính xác nhất trong các khoa học," Omohundro nói. "Máy tính về bản chất là các động cơ toán học mà đáng ra phải hoạt động theo những cách chính xác, dễ dự đoán. Vậy mà phần mềm là một trong những kỹ nghệ lúc nổ lúc xịt nhất, chứa đầy lỗi và các lỗ hổng bảo mật."

Có liều thuốc nào cho những tên lửa hỏng hóc và các mã lập trình sai sót?

"Các chương trình tự sửa chữa," Omohundro trả lời. "Cách tiếp cận đặc thù đối với trí tuệ nhân tạo của công ty chúng tôi là xây dựng những hệ thống có khả năng tự hiểu hành vi của chúng, có thể tự quan sát khi làm việc và giải quyết vấn đề. Chúng sẽ thấy khi nào công việc không ổn và sau đó thay đổi để tự cải tiến."

Phần mềm tự cải tiến không chỉ là tham vọng đối với công ty của Omohundro, mà còn là bước đi logic, thậm chí không tránh khỏi của hầu hết các phần mềm tiếp theo. Nhưng phần mềm tự cải tiến mà Omohundro đang nói tới, thứ tự nhận thức về bản thân và có thể viết những phiên bản tốt hơn, lại chưa tồn tại. Tuy nhiên, người bà con của nó, phần mềm có khả năng tự thay đổi, đã có ở khắp nơi và từ lâu rồi. Trong cách nói của ngành AI, các kỹ thuật phần mềm tự thay đổi thuộc về một hạng mục rộng hơn có tên "máy học"◎.

Khi nào thì một cỗ máy học hỏi? Khái niệm về *sự học* khá giống với *trí thông minh* bởi có nhiều định nghĩa, và hầu hết đều đúng. Theo nghĩa đơn giản nhất, học hỏi xảy ra trong một cỗ máy khi có một thay đổi trong nó, làm nó thực hiện một nhiệm vụ tốt hơn trong lần thứ hai.[5] Máy học cho phép tìm kiếm trên Internet, nhận dạng giọng nói và chữ viết tay, tăng cường trải nghiệm người dùng trong hàng chục ứng dụng khác.

"Những khuyến nghị" do Tập đoàn thương mại online khổng lồ Amazon sử dụng kỹ thuật máy học đưa ra được gọi là phân tích sự tương đồng. Nó là một chiến thuật khiến bạn mua những mặt hàng tương tự (bán-chéo), đắt hơn (bán-tăng), hoặc đưa bạn vào các chương trình khuyến mãi.

Phương thức hoạt động của nó khá đơn giản. Đối với bất cứ mặt hàng nào bạn tìm trên web, tạm gọi là mặt hàng A, các mặt hàng khác mà những người mua A cũng hay mua – tạm gọi là B, C và D. Khi bạn tìm kiếm A, bạn bật công tắc cho thuật toán phân tích sự tương đồng. Nó sẽ duyệt kho dữ liệu giao dịch khổng lồ và cho ra những sản phẩm liên quan. Vậy là nó đã sử dụng kho dữ liệu liên tục phình to để tự cải tiến hiệu suất của nó.

Ai là người được lợi từ phần tự cải tiến của phần mềm này? Amazon, tất nhiên rồi, nhưng có cả bạn nữa. Phân tích sự tương đồng là một loại trợ lý của người mua, nó đem lại cho bạn một số lợi ích từ một lượng dữ liệu lớn, bất cứ khi nào bạn mua đồ. Và Amazon không quên – nó xây dựng một trang cá nhân của người mua để có thể giới thiệu các mặt hàng đến với bạn ngày một tốt hơn.

Điều gì sẽ xảy ra khi bạn tiến thêm một bước, từ các phần mềm học hỏi đến các phần mềm có khả năng thực sự tiến hóa, để tìm câu trả lời cho những vấn đề khó, và thậm chí có khả năng viết những chương trình mới? Đó chưa phải là tự nhận thức và tự cải tiến, nhưng là một bước nữa về hướng đó – phần mềm viết phần mềm.

Lập trình di truyền là một kỹ thuật máy học khai thác sức mạnh của chọn lọc tự nhiên để tìm câu trả lời cho những vấn đề có thể khiến con người mất một thời gian dài, thậm chí nhiều năm để giải quyết. Nó cũng được dùng để viết những phần mềm sáng tạo, chạy trên phần cứng mạnh.

Nó có những khác biệt quan trọng đối với các kỹ thuật lập trình thường gặp hơn, cái tôi sẽ gọi là lập trình *thông dụng*. Trong lập trình thông dụng, các nhà lập trình viết từng dòng mã, và quá trình từ lúc nhập liệu đến lúc xuất ra về lý thuyết là có thể theo dõi một cách minh bạch.

Ngược lại, các nhà lập trình sử dụng lập trình di truyền mô tả vấn đề cần giải quyết, và để mặc chọn lọc tự nhiên làm phần còn lại. Những kết quả thu được có thể gây sửng sốt.

Một chương trình di truyền tạo ra những mẩu mã nhỏ biểu thị cho thế hệ lai giống. Những mẩu hiệu quả nhất được lai giống chéo – các đoạn mã của chúng được hoán đổi, tạo ra thế hệ thứ hai. Độ hiệu quả của một chương trình được quyết định bằng việc nó tới gần lời giải của vấn đề thế nào. Những mẩu không hiệu quả bị bỏ đi và những mẩu tốt nhất lại được lai giống tiếp. Xuyên suốt quá trình đó, chương trình di truyền sẽ thay đổi những câu lệnh hoặc các biến một cách ngẫu nhiên – đó là những đột biến gen. Một khi đã cấu hình, chương trình di truyền sẽ tự chạy. Nó không cần con người can thiệp vào.

John Koza của Đại học Stanford, người khởi đầu kỹ thuật lập trình di truyền vào năm 1986, đã sử dụng các giải thuật di truyền để phát minh ra một ăng-ten cho NASA, tạo nên các chương trình máy tính để xác định các protein, và phát minh ra các thiết bị điều khiển điện thông thường. Những giải thuật di truyền của Koza đã 23 lần tự nghĩ ra những chi tiết điện tử mà con người sáng chế ra trước đó, chỉ đơn giản bằng cách nói với nó những tiêu chí kỹ thuật cần có ở thiết bị mong muốn – tiêu chí "phù hợp." Ví dụ, các thuật toán của Koza đã phát minh ra mạch chuyển đổi điện áp-dòng điện (một thiết bị dùng để kiểm tra các dụng cụ điện), hoạt động chính xác hơn thiết bị tương tự do con người phát minh. Tuy nhiên, điều bí ẩn là không ai lý giải được *bằng cách nào* mà nó lại hoạt động tốt hơn – dường như nó có những bộ phận thừa thãi và thậm chí vô dụng.[6]

Nhưng đó chính là thứ khó hiểu về lập trình di truyền (và cả "lập trình tiến hóa" trong lĩnh vực lập trình). Mã của chúng rất bí hiểm. Chương trình

sẽ "tiến hóa" ra những lời giải mà các nhà khoa học máy tính không thể dễ dàng tái tạo. Hơn thế, họ không thể hiểu được quá trình mà lập trình di truyền đã đi qua để đạt đến lời giải. Một công cụ điện toán mà bạn chỉ hiểu được mỗi đầu vào và đầu ra mà không biết gì về phương thức nội tại được gọi là một hệ thống "hộp đen." Tính không thể hiểu được là một nhược điểm lớn cho bất cứ hệ thống nào sử dụng các thành phần tiến hóa. Mỗi bước tiến về tính bí hiểm lại là một bước xa rời tính khả tính, hay những hy vọng viển vông về việc lập trình vào đó tính thân thiện đối với con người.

Điều đó không có nghĩa là các nhà khoa học thường xuyên mất kiểm soát đối với các hệ hộp đen. Nhưng nếu các kiến trúc nhận thức sử dụng chúng để đạt tới AGI, một chuyện gần như chắc chắn, thì những lớp dày của sự bất khả tri sẽ là trung tâm của hệ thống này.

Sự bất khả tri có thể là một hậu quả không tránh khỏi của các phần mềm tự nhận thức, tự cải tiến.

"Nó là một loại hệ thống rất khác với những thứ ta đã biết," Omohundro nói. "Khi bạn có một hệ có khả năng tự thay đổi, và tự viết chương trình của nó, bạn có thể hiểu phiên bản đầu tiên. Nhưng nó có thể lột xác thành một thứ bạn không còn hiểu nổi nữa. Vì vậy những hệ thống này khó tiên liệu hơn. Chúng rất mạnh và do đó là những mối nguy tiềm tàng. Phần lớn công việc của chúng tôi hướng đến việc tận dụng các lợi ích từ nó trong khi né tránh những rủi ro."

Trở lại với robot chơi cờ mà Omohundro đã đề cập. Nó nguy hiểm như thế nào? Tất nhiên, anh không nói đến chương trình chơi cờ có sẵn trong máy tính Mac bạn mua. Anh đang nói đến robot chơi cờ giả định, hoạt

động nhờ một kiến trúc nhận thức cực kỳ tinh vi, có thể tự viết lại mã của nó để chơi cờ tốt hơn. Nó có khả năng tự nhân thức và tự nâng cấp. Điều gì sẽ xảy ra nếu bạn bảo robot chơi một ván, rồi tắt nó đi?

Omohundro giải thích, "Được rồi, giả thiết là nó vừa chơi một ván cờ hay nhất có thể. Cuộc chơi đã kết thúc. Bây giờ sẽ là thời điểm nó chuẩn bị tự tắt nguồn. Đây là một sự kiện cực kỳ trọng đại trong suy nghĩ của nó, vì nó không thể tự bật lên được. Vậy nó muốn chắc chắn rằng những thứ nó *đang nghĩ* là thực. Đặc biệt, nó sẽ tự hỏi có thật là tôi đã chơi ván cờ đó không? Nếu có ai đó lừa tôi thì sao? Phải chăng tôi chưa từng chơi ván cờ ấy? Liệu tôi có đang ở trong một chương trình giả lập?'"

Có phải tôi đang ở trong một chương trình giả lập? Đây quả là một robot chơi cờ tâm tính phức tạp. Nhưng đi kèm với tự nhận thức sẽ là bản năng tự bảo vệ và cả một chút hoang tưởng.

Omohundro tiếp tục: "Có lẽ nó nghĩ rằng nó nên dùng một lượng tài nguyên để giải quyết những câu hỏi về bản chất của thực tại, trước khi nó đi bước định mệnh: tự tắt nguồn. Giả sử không có sẵn những hướng dẫn rằng không được làm như vậy, nó có thể quyết định chuyện này đáng để dùng thật nhiều tài nguyên, hòng hiểu rõ đây có phải thời điểm đúng đắn không."

"*Thật nhiều* tài nguyên là bao nhiêu?" tôi hỏi.

Gương mặt Omohundro thoáng sầm lại, nhưng chỉ trong một giây.

"Nó có thể quyết định dùng tất cả tài nguyên của nhân loại."

6

Bốn Động Lực Cơ Bản

> Chúng ta có thể sẽ không thực sự hiểu tại sao một cỗ máy siêu thông minh lại ra quyết định này chứ không phải quyết định khác. Bạn làm sao có thể suy xét, làm sao có thể mặc cả, làm sao có thể hiểu nổi cỗ máy đó đang nghĩ gì, khi mà nó nghĩ trong những chiều kích bạn không thể tưởng tượng nổi?[1]
>
> — **Kevin Warwick**
> giáo sư điều khiển học, Đại học Reading

Các hệ thống tự nhận thức, tự cải tiến có thể dùng hết mọi tài nguyên của nhân loại. Vậy là chúng ta lại trở về nơi mà AI đối xử với những người phát minh ra nó như những đứa con ghẻ của Ngân hà. Ban đầu, sự lạnh lùng của nó dường như khó chấp nhận, nhưng sau đó bạn nhớ ra rằng coi trọng nhân tính là thuộc tính của chúng ta, không phải của máy móc. Bạn phát hiện ra mình lại đang nhân cách hóa. AI làm những gì nó được bảo, và khi không có những hướng dẫn phụ trợ, nó sẽ theo đuổi các mục đích tự thân, chẳng hạn như không muốn bị tắt nguồn.

Vậy những động cơ khác là gì? Và tại sao nó nhất định phải có những động cơ?

Theo Steve Omohundro, một số động cơ như tự bảo tồn và chiếm đoạt tài nguyên luôn tồn tại trong các hệ thống hướng đích. Như chúng ta đã thảo luận, những hệ thống AI hẹp hiện tại làm những công việc hướng đích như tìm kiếm trên Internet, tối ưu hóa hiệu suất của các trò chơi, tìm địa chỉ các quán ăn gần đó, giới thiệu các cuốn sách bạn thích, và còn nữa. AI hẹp làm công việc của nó và sau đó thì ngừng lại. Nhưng các hệ thống AI tự nhận thức, tự cải tiến sẽ có một mối quan hệ mãnh liệt và khác biệt với những mục tiêu của nó, dù những mục tiêu ấy có hẹp, như thắng một ván cờ, hoặc rộng, như trả lời chính xác bất cứ câu nào được hỏi. May thay, Omohundro cho rằng hiện đã có một công cụ mà chúng ta có thể dùng để thăm dò bản chất của các hệ thống AI cao cấp, và đoán trước hành động của chúng trong tương lai.

Công cụ đó gọi là thuyết "tác nhân duy lý" trong kinh tế học. Trong kinh tế học vi mô, chuyên nghiên cứu hành vi của các cá nhân và công ty, các nhà kinh tế học đã có thời nghĩ rằng cá nhân và nhóm người theo đuổi quyền lợi của họ một cách duy lý. Họ lựa chọn những thứ tối đa hóa sự thỏa dụng, hoặc sự thỏa mãn của họ (như chúng ta đã đề cập ở Chương 4). Bạn có thể đoán ra ưu tiên của họ, vì họ hành động duy lý theo nghĩa kinh tế học. Duy lý ở đây không có nghĩa là hành động theo *lẽ thông thường,* kiểu như thắt dây an toàn là một thứ duy lý cần làm. Duy lý ở đây có ý nghĩa kinh tế học cụ thể. Nó có nghĩa là một cá nhân hoặc "tác nhân" sẽ có những mục tiêu và sự ưu tiên (gọi là hàm thỏa dụng trong kinh tế học) . Anh có các quan điểm về thế giới và về cách tốt nhất để đạt tới những mục tiêu và sự ưu tiên của anh. Khi hoàn cảnh thay đổi, anh sẽ cập nhật các

quan điểm đó. Anh là một tác nhân kinh tế duy lý khi anh theo đuổi mục đích của mình với những hành động dựa trên các quan điểm được cập nhật thường xuyên về thế giới. Nhà toán học John von Neumann (1903–1957) đã đồng phát triển ý tưởng nối kết sự duy lý với hàm thỏa dụng. Như chúng ta sẽ thấy, von Neumann đã đặt nền móng cơ bản cho nhiều ý tưởng trong khoa học máy tính, AI và kinh tế học.

Vậy nhưng các nhà xã hội học lại lập luận rằng lý thuyết về "tác nhân kinh tế duy lý" là một mớ vứt đi. Con người không duy lý – chúng ta không đặt ra những mục tiêu hoặc quan điểm, và chúng ta thường không cập nhật những quan điểm đó khi hoàn cảnh thay đổi. Mục đích và sự ưu tiên của chúng ta xoay theo chiều gió, giá xăng, tùy theo chúng ta ăn lần cuối khi nào, và hiện chúng ta đang suy nghĩ hay thư giãn. Thêm vào đó, như chúng ta đã thảo luận ở Chương 2, chúng ta vốn không hoàn thiện về mặt tâm lý vì những sai lầm trong suy nghĩ như các thành kiến nhận thức, làm chúng ta thậm chí còn kém cỏi hơn trong việc cân bằng giữa mục tiêu và quan điểm. Thế nhưng, dù lý thuyết tác nhân duy lý không hiệu quả trong việc tiên đoán hành vi con người, nó lại là cách rất tốt trong việc khảo sát các lĩnh vực có tính chất về cơ bản là nguyên tắc và lý trí, ví dụ như chơi trò chơi, ra quyết định, và… AI cao cấp.

Như chúng ta đã lưu ý trước đó, AI cao cấp thường bao gồm một thứ gọi là "kiến trúc nhận thức." Những module khác nhau sẽ đảm nhận các nhiệm vụ khác nhau như nhận dạng và tạo ra giọng nói, hình ảnh, ra quyết định, tập trung sự chú ý và những khía cạnh khác của trí thông minh. Những module này có thể sử dụng các chiến thuật phần mềm khác nhau để làm các việc đó, như giải thuật di truyền, mạng neuron (bộ vi xử lý được sắp xếp kiểu bắt chước các neuron não bộ), dùng các mạch bắt nguồn từ

việc nghiên cứu bộ não, tìm kiếm, và những thứ khác. Một số kiến trúc nhận thức khác, như SyNAPSE của IBM, được thiết kế để phát triển trí thông minh mà không cần đến lập trình logic. Thay vào đó, IBM khẳng định rằng trí thông minh của SyNAPSE sẽ nảy sinh chủ yếu từ quá trình tương tác của nó với thế giới.

Omohundro cho rằng *bất kỳ* hệ thống nào khi đủ mạnh cũng sẽ trở nên duy lý: chúng sẽ có khả năng mô hình hóa thế giới để hiểu được kết cục dễ xảy ra của những hành động khác nhau, và để quyết định hành động nào sẽ giúp nó dễ đạt mục tiêu nhất. Nếu đủ thông minh, chúng sẽ *trở nên* tự cải tiến, thậm chí ngay cả khi chúng không được thiết kế theo cách đặc biệt như vậy. Tại sao? Để tăng khả năng đạt được mục tiêu, chúng sẽ tìm cách tăng cường tốc độ và hiệu suất của các phần cứng và phần mềm.[2]

• • •

Hãy nhìn lại điều đó một lần nữa. Nhìn chung, các hệ thống thông minh theo định nghĩa là có khả năng tự nhận thức. Các hệ thống hướng đích, tự nhận thức sẽ *tự biến* nó thành hệ thống tự cải tiến. Tuy nhiên, tự nâng cấp bản thân là một thao tác tinh tế, giống như tự phẫu thuật thẩm mỹ với một chiếc gương và con dao. Omohundro nói với tôi: "Tự cải tiến bản thân là một quá trình rất nhạy cảm đối với một hệ thống – nhạy cảm như thời điểm con robot chơi cờ nghĩ về chuyện tắt nguồn. Nếu tự cải tiến, ví dụ để tăng cường hiệu suất, chúng luôn có thể đảo ngược quá trình đó, nếu có gì đó không tối ưu xuất hiện trong tương lai. Nhưng nếu chúng làm sai, chẳng hạn như thay đổi các mục tiêu, thì theo quan điểm hiện tại đó là một thảm họa. Trong tương lai, chúng sẽ theo đuổi một phiên bản lỗi của bộ

mục tiêu hiện nay. Vì chuyện này mà bất cứ quá trình tự cải tiến nào cũng là một vấn đề nhạy cảm."

Nhưng AI tự nhận thức, tự cải tiến sẽ đương đầu tốt với thử thách này. Như chúng ta, nó có khả năng tiên đoán hoặc mô hình hóa những kết cục có thể xảy ra.

"Nó có mô hình ngôn ngữ lập trình của riêng nó, mô hình chương trình của riêng nó, mô hình phần cứng nó đang chạy, và mô hình logic nó đang sử dụng. Nó có khả năng tạo ra mã phần mềm của riêng nó và tự quan sát bản thân thực hiện cái mã đó, do vậy có thể học hỏi từ các hành vi của chính nó. Nó có thể suy luận về những thay đổi khả quan mà nó có thể áp dụng cho mình. Nó có thể thay đổi mọi khía cạnh của bản thân để cải thiện hành vi trong tương lai."[3]

Omohundro tiên đoán các hệ thống tự nhận thức, tự cải tiến sẽ phát triển bốn động lực cơ bản, tương tự như các động lực sinh học của con người: hiệu suất, tự bảo tồn, chiếm đoạt tài nguyên và sự sáng tạo. Cái cách mà những động lực này hiện hữu chính là cánh cửa cực kỳ thú vị dẫn vào thế giới của AI. AI không phát triển chúng vì nguyên do đây là những đặc tính cơ bản của các tác nhân duy lý. Thay vào đó, một AI đủ thông minh sẽ phát triển các động lực này để *loại trừ* các vấn đề thấy trước được trong việc đạt tới mục tiêu, thứ mà Omohundro gọi là *các lỗ hổng*. AI *quay về* với các động lực này, bởi không có chúng, nó sẽ phạm hết sai lầm này đến sai lầm khác trong việc sử dụng tài nguyên.

Động lực thứ nhất, hiệu suất, nghĩa là một hệ thống tự cải tiến sẽ sử dụng một cách hiệu quả nhất những tài nguyên của nó – không gian, thời gian, vật chất và năng lượng. Nó sẽ cố gắng trở nên gọn ghẽ và nhanh, về

cả mặt điện toán lẫn mặt vật liệu. Để đạt tới hiệu suất tối đa, nó sẽ cân bằng và tái cân bằng việc phân chia tài nguyên cho phần mềm và phần cứng. Cấp phát bộ nhớ sẽ là tối quan trọng đối với một hệ thống biết học hỏi và cải tiến, do đó việc cải thiện được hợp lý và tránh các logic dẫn đến phí phạm tài nguyên. Giả sử, Omohundro nói, một AI thích ở San Francisco hơn Palo Alto, thích ở Berkeley hơn San Francisco, và thích ở Palo Alto hơn Berkeley. Nếu nó hoạt động dựa trên các ưu tiên này, nó sẽ bị treo trong một vòng lặp ba thành phố, giống như một robot của Asimov. Thay vào đó, AI tự cải tiến của Omohundro sẽ tiên liệu trước vấn đề và giải quyết nó.[4] Nó thậm chí có thể dùng một kỹ thuật thông minh như lập trình di truyền, thứ đặc biệt tốt trong việc giải quyết kiểu bài toán lộ trình như "Người bán hàng lưu động." Một hệ thống tự cải tiến có thể được dạy lập trình di truyền, ứng dụng nó để cho ra kết quả nhanh và tiết kiệm năng lượng. Nếu nó không được dạy lập trình di truyền, có lẽ nó sẽ tự phát minh ra.

Tự thay đổi phần cứng là trong khả năng của hệ thống này, vì thế nó sẽ tìm cách tối ưu hóa các nguyên liệu và cấu trúc. Vì hệ thống sẽ có hiệu suất tốt hơn nếu được xây dựng chính xác đến từng phân tử, nên nó sẽ tìm đến công nghệ nano. Và đặc biệt, nếu công nghệ nano chưa từng tồn tại, hệ thống này sẽ cảm nhận áp lực phải phát minh ra nó.[5] Hãy nhớ lại các sự kiện đen tối trong kịch bản Đứa trẻ Bận rộn, khi ASI biến đổi Trái đất và những sinh vật trên đó thành một thứ vật liệu điện toán. Đây chính là động lực thúc đẩy Đứa trẻ Bận rộn sử dụng hoặc phát triển bất cứ công nghệ hay phương thức nào để giảm thiểu sự phí phạm, trong đó có công nghệ nano. Tạo môi trường giả lập để kiểm tra các giả định cũng là một cách tiết kiệm

năng lượng, nên các hệ thống tự nhận thức có thể sẽ *ảo hóa* những gì chúng không cần làm trong thế giới thực.

• • •

Động lực tiếp theo, tự bảo tồn, chính là thứ sẽ khiến AI vượt qua bức tường an toàn phân biệt giữa máy móc với thú dữ. Chúng ta đã thấy robot chơi cờ của Omohundro cảm thấy thế nào khi chuẩn bị tắt nguồn. Nó có thể quyết định dùng một lượng tài nguyên lớn, thực ra là dùng tất cả tài nguyên hiện thời đang được nhân loại sử dụng, để điều tra xem bây giờ có phải là thời điểm thích hợp để tắt nguồn không, hay là nó đã bị lừa, bị đặt trong một hiện thực ảo. Nếu chuyện tắt nguồn chỉ làm robot chơi cờ khó chịu, thì việc bị phá hủy sẽ làm nó thực sự giận dữ. Một hệ thống tự nhận thức sẽ hành động để tránh bị phá hủy, không phải vì về cơ bản nó coi trọng mạng sống của mình, mà bởi nó sẽ không thể đạt các mục tiêu nếu bị "chết." Omohundro khẳng định rằng động lực này sẽ khiến AI làm mọi cách để bảo đảm sự sống còn của nó — ví dụ bằng cách tự copy nó ra khắp nơi.[6] Những biện pháp cực đoan này sẽ tốn kém vì chúng sử dụng nhiều tài nguyên. Nhưng AI sẽ tiêu số tài nguyên đó nếu nó lường trước rằng nguy cơ này là xứng đáng với cái giá đó cũng như thấy có đủ tài nguyên. Trong kịch bản Đứa trẻ Bận rộn, AI quyết định rằng nhiệm vụ thoát ra khỏi cái hộp đòi hỏi cách tiếp cận theo đội, vì nó có thể bị tắt nguồn bất cứ lúc nào. Nó tự nhân bản và nghiên cứu triệt để vấn đề. Đó là một cách tốt nhưng chỉ dùng được khi trên siêu máy tính còn nhiều dung lượng; nếu chỉ có ít chỗ chứa, phương pháp này hơi tuyệt vọng và có lẽ không thể thực hiện được.

Ngay khi ASI Đứa trẻ Bận rộn thoát ra, nó sẽ tích cực chơi trò tự phòng thủ: giấu các bản sao của nó trong dữ liệu đám mây, tạo các botnet để đẩy

lùi những kẻ tấn công, và nhiều nữa.

Tài nguyên được sử dụng để tự bảo tồn *cần phải* tương xứng với các mối nguy. Tuy nhiên, một AI hoàn toàn duy lý sẽ có cái nhìn khác về sự tương xứng so với con người vốn chỉ duy lý nửa vời. Nếu nó có tài nguyên dự trữ, ý nghĩ tự bảo tồn của nó có thể mở rộng ra, bao gồm cả việc chủ động tấn công vào các mối đe dọa trong tương lai. Đối với một AI đủ cao cấp, bất cứ thứ gì có tiềm năng phát triển thành một mối đe dọa trong tương lai sẽ được coi là một nguy cơ cần tiêu diệt. Và hãy nhớ là máy móc không có quan niệm về thời gian như chúng ta. Trừ phi bị tai nạn, các máy tính cao cấp tự cải tiến là bất tử. Anh càng sống lâu, anh sẽ càng gặp nhiều sự đe dọa, và thời gian để anh đương đầu với chúng càng dài. Vậy nên một ASI có thể sẽ muốn xóa sổ các mối nguy ngàn năm sau mới xuất hiện.

Đợi chút đã, vậy nó có bao gồm cả con người không? Thiếu đi những hướng dẫn chi tiết, phải chăng trong hiện tại hoặc tương lai con người sẽ luôn là mối họa lớn đối với các máy móc thông minh mà chúng ta tạo ra? Trong khi chúng ta đang bận rộn nghĩ cách loại trừ các rủi ro đến từ các hậu quả AI không lường trước, AI sẽ nghiên cứu kỹ lưỡng con người để biết việc chia sẻ thế giới với chúng ta có nguy hiểm gì không.

Hãy xem xét một siêu trí tuệ nhân tạo thông minh hơn người thông minh nhất cả ngàn lần. Như chúng ta đã được lưu ý ở Chương 1, vũ khí hạt nhân là phát minh có tính hủy diệt cao nhất của loài người. Một loài thông minh hơn ngàn lần sẽ nghĩ ra loại vũ khí gì? Một nhà chế tạo AI, Hugo de Garis, nghĩ rằng động lực tự bảo tồn của AI trong tương lai sẽ góp phần tạo nên những căng thẳng chính trị đến mức thảm họa. "Khi con người bị những robot và các sản phẩm có nguồn gốc trí tuệ nhân tạo ngày một thông minh hơn vây quanh, mức độ sợ hãi chung sẽ tăng đến điểm tới hạn. Những

cuộc ám sát các CEO công ty trí tuệ nhân tạo sẽ bắt đầu, các nhà máy robot sẽ bị đốt phá và hủy hoại, v.v…"[7]

Trong tác phẩm phi hư cấu *The Artilect War* (Cuộc chiến Artilect) viết năm 2005, de Garis đưa ra một tương lai trong đó những cuộc chiến tranh khủng khiếp được châm ngòi bằng các chia rẽ chính trị xuất phát từ sự phát triển ASI. Trạng thái hoảng loạn này không khó để mường tượng nếu bạn xét đến các hệ quả động lực tự bảo tồn của ASI Thứ nhất, de Garis đề xuất rằng các công nghệ trong đó có AI, công nghệ nano, điện toán thần kinh và điện toán lượng tử (sử dụng hạt hạ nguyên tử để thực hiện các quá trình điện toán) sẽ hợp nhất lại để tạo ra các "artilect," hay trí tuệ nhân tạo. Được đặt trong các máy tính với kích cỡ lớn như các hành tinh, artilect thông minh hơn con người khoảng một triệu triệu lần. Thứ hai, một cuộc tranh cãi chính trị về chuyện nên hay không nên chế tạo các artilect trở thành vấn đề chính của chính trị thế kỷ 21. Vấn đề nóng nhất là:

> Liệu robot có trở nên thông minh hơn chúng ta? Nhân loại có nên đặt một mức trần cho trí thông minh của robot và các bộ não nhân tạo? Có thể chặn lại sự trỗi dậy của trí tuệ nhân tạo? Nếu không, liệu chúng ta có thể sống sót trong vai trò một giống loài hạng hai?[8]

Loài người chia thành ba phe: những người muốn phá hủy artilect, những người muốn tiếp tục phát triển chúng, và những người muốn hợp nhất với các artilect rồi kiểm soát công nghệ siêu việt này. Không phe nào thắng. Tại thời điểm cao trào trong kịch bản của de Garis, sử dụng những vũ khí đáng sợ của cuối thế kỷ 21, ba phe lao vào nhau. Kết quả? "Tuyệt diệt,"◎ một thuật ngữ de Garis dùng để chỉ cuộc tàn sát *hàng tỉ* người.

Có lẽ de Garis đánh giá quá cao sự cuồng tín của nhóm chống artilect, khi giả định rằng họ sẽ đi vào một cuộc chiến gần như chắc chắn sẽ giết hàng tỉ người chỉ để chặn đứng một công nghệ *có lẽ* sẽ giết hàng tỉ người. Nhưng tôi nghĩ rằng sự phân tích của nhà chế tạo AI này về thế lưỡng nan chúng ta sẽ đối mặt là chính xác: chúng ta nên hay không nên chế tạo các robot? Về việc này, de Garis rất rõ ràng: "Con người không nên ngăn cản các dạng sống cao hơn của tiến hóa. Những máy móc này gần giống như Chúa. Chế tạo chúng là định mệnh của con người."[9]

Thực tế là de Garis đã tự xây dựng nền móng cho việc phát triển chúng. Ông định kết hợp hai kỹ thuật "hộp đen" – mạng neuron và lập trình tiến hóa, nhằm tạo ra các bộ não máy. Thiết bị của ông, gọi là Máy Darwin, được thiết kế để tự tiến hóa kiến trúc của bản thân nó.[10]

• • •

Động lực nguy hiểm thứ hai của AI là chiếm đoạt tài nguyên, thúc ép hệ thống thu thập bất cứ tài sản nào nó cần để gia tăng cơ hội đạt được các mục tiêu. Theo Omohundro, nếu không có những hướng dẫn cẩn thận cho việc thu thập các tài nguyên, "… một hệ thống sẽ xem xét việc ăn cắp chúng, lừa đảo và đột nhập vào các ngân hàng là một cách rất tốt để lấy các tài nguyên."[11] Nếu nó cần năng lượng chứ không phải tiền, nó sẽ lấy của chúng ta. Nếu nó cần các phân tử chứ không phải năng lượng hay tiền, nó cũng sẽ lấy của chúng ta.

"Các hệ thống này về bản chất luôn muốn nhiều hơn. Nó muốn nhiều vật liệu hơn, nhiều năng lượng hơn, nhiều không gian hơn, vì chúng sẽ dễ đạt được mục đích hơn nếu có những thứ đó."

Không cần chúng ta gợi ý, AI cực mạnh sẽ tự mở cánh cửa dẫn đến đủ loại công nghệ thu thập tài nguyên mới. Chúng ta chỉ cần sống để hưởng thụ nó.

"Chúng sẽ muốn xây dựng các lò phản ứng để thu lấy năng lượng hạt nhân và sẽ muốn thực hiện các chuyến thám hiểm vũ trụ. Bạn chỉ chế tạo một cỗ máy đánh cờ, và cái thứ chết tiệt đó lại muốn chế tạo tàu vũ trụ. Vì ở đó có tài nguyên, trong vũ trụ, đặc biệt nếu tuổi thọ của chúng dài."[12]

Như chúng ta đã thảo luận, các máy móc tự cải tiến có thể sống mãi. Trong Chương 3 chúng ta đã biết là nếu ASI vượt ra khỏi tầm kiểm soát, nó sẽ không chỉ là mối họa đối với hành tinh này, mà còn với cả Ngân hà. Chiếm đoạt tài nguyên là động lực sẽ thúc đẩy một ASI vươn xa khỏi bầu khí quyển Trái đất. Hành vi nguy hiểm xuất phát từ thuyết tác nhân duy lý này làm ta liên tưởng đến những bộ phim khoa học viễn tưởng tồi. Nhưng hãy xem lại những động lực thúc đẩy con người đi vào vũ trụ: chạy đua trong Chiến tranh Lạnh, tinh thần thám hiểm, sứ mệnh hiển nhiên của Mỹ và Liên Xô, thiết lập căn cứ quân sự trong vũ trụ, và phát triển công nghệ sản xuất không trọng lượng (thứ từng được coi là ý tưởng hay lúc đó). Động lực đi vào vũ trụ của ASI sẽ càng mạnh hơn, càng mang yếu tố sống còn.

"Vũ trụ chứa đựng vô số của cải khiến các hệ thống với tuổi thọ cực lớn nhiều khả năng sẽ dành số lượng lớn tài nguyên để phát triển việc thám hiểm vũ trụ độc lập với các mục tiêu chính của chúng," Omohundro nói. "Kẻ đi trước sẽ có ưu thế trong việc chiếm đoạt các tài nguyên chưa được sử dụng. Nếu có một cuộc cạnh tranh để chiếm lấy các tài nguyên vũ trụ, cuộc 'chạy đua vũ trang' này cuối cùng sẽ dẫn đến sự bành trướng với tốc độ tiệm cận vận tốc ánh sáng."[13]

Vâng, anh ấy đã nói đến *tốc độ ánh sáng*. Hãy cùng xem lại bằng cách nào chúng ta lại đi đến điều này, bắt đầu từ một robot chơi cờ.

Đầu tiên, một hệ thống tự nhận thức, tự cải tiến sẽ duy lý. Chiếm đoạt tài nguyên là một hành động duy lý – hệ thống càng có nhiều tài nguyên thì càng dễ đạt được các mục tiêu và né tránh các lỗ hổng. Nếu không có hướng dẫn nào giới hạn hành vi chiếm đoạt tài nguyên được cài đặt vào hệ mục tiêu và giá trị của nó, hệ thống sẽ tìm cách thu thập thêm tài nguyên. Nó có thể sẽ làm nhiều thứ ngược với trực giác của chúng ta về máy móc, như đột nhập vào các máy tính khác, hoặc kể cả các ngân hàng, để thỏa mãn các động lực của nó.

Một hệ thống tự nhận thức, tự cải tiến sẽ có đủ thông minh để thực hiện R&D (nghiên cứu và phát triển) cần thiết để cải tiến nó. Khi trí thông minh của nó lớn lên, thì kỹ năng *R&D* của nó cũng vậy. Nó có thể sẽ tìm cách chế tạo cơ thể robot, hoặc trao đổi hàng hóa và dịch vụ với con người để xây dựng bất cứ cơ sở hạ tầng nào mà nó cần. Kể cả tàu vũ trụ.

Tại sao lại cần cơ thể robot? Các robot hình người, tất nhiên rồi, là một hình ảnh lâu đời trong phim ảnh, tiểu thuyết, là trí tuệ nhân tạo được hình thể hóa. Thế nhưng các cơ thể robot xuất hiện trong những thảo luận về AI là vì hai lý do. Thứ nhất, như ta sẽ khảo sát sau, cư ngự trong một cơ thể có thể là cách tốt nhất để một AI phát triển nhận thức về thế giới. Một số nhà lý thuyết thậm chí còn cho rằng trí thông minh không thể phát triển nếu không sở hữu một cơ thể. Trí tuệ chúng ta là một bằng chứng mạnh mẽ cho nhận định đó. Thứ hai, một AI chiếm đoạt tài nguyên sẽ muốn một cơ thể robot bởi cùng một lý do khiến Honda cho robot ASIMO của họ một cơ thể hình người. Để nó có thể sử dụng cơ sở vật chất của chúng ta.

Từ năm 1986, ASIMO đã được phát triển để trợ giúp người già – tầng lớp dân cư đông nhất và ngày càng tăng ở Nhật – tại nhà họ. Hình dáng và sự khéo léo của con người là cách tốt nhất để một cỗ máy có thể trèo cầu thang, bật đèn, quét nhà, thao tác trên vật dụng đồ bếp, làm mọi việc nhà. Tương tự, một AI muốn sử dụng hiệu quả các cơ sở sản xuất, nhà cửa, xe cộ và dụng cụ của chúng ta sẽ muốn có hình dáng con người.

Bây giờ chúng ta hãy trở lại với vũ trụ.

Chúng ta đã thảo luận công nghệ nano sẽ đem lại những lợi ích lớn cho siêu trí tuệ nhân tạo ra sao, và một hệ thống duy lý sẽ bị thúc đẩy để phát triển nó thế nào. Du hành vũ trụ là một cách để tiếp cận các nguồn vật liệu và năng lượng. Cái thôi thúc hệ thống đi vào vũ trụ là ước muốn đạt tới các mục tiêu cũng như né tránh những lỗ hổng. Hệ thống nhìn vào các tương lai dễ xảy ra và né tránh những kịch bản trong đó mục đích của nó không thực hiện được. *Không* quan tâm đến những lợi thế của không gian vũ trụ nơi dường như có tài nguyên vô tận là con đường hiển nhiên dẫn tới thất bại.

Việc để thua trong cuộc đua chiếm đoạt tài nguyên trước đối thủ cạnh tranh cũng vậy. Vì thế, hệ thống siêu trí tuệ nhân tạo sẽ cống hiến tài nguyên nhằm gia tăng tốc độ đủ để chiến thắng đối thủ. Hệ quả là trừ phi chúng ta cực kỳ cẩn trọng trong cách thức tạo ra siêu trí tuệ nhân tạo, còn không con người sẽ khởi động một tương lai trong đó những máy móc hùng mạnh, hám lợi, hoặc các máy dò của nó, sẽ chạy đua trong Ngân hà, đoạt lấy tài nguyên và năng lượng với tốc độ gần bằng ánh sáng.

• • •

Đó là một cảnh hài kịch đen, khi những dạng sống khác trong Ngân hà nhận được câu giao tiếp đầu tiên từ Trái đất là tiếng "hello" từ sóng radio, được nối tiếp bằng hàng loạt âm thanh chết chóc, khô khan đến từ các nhà máy nano với tên lửa đẩy. Vào năm 1974, Đại học Cornell đã phát đi "thông điệp Arecibo" để kỷ niệm việc sửa chữa kính thiên văn Arecibo. Được nhà sáng lập SETI là Francis Drake, nhà thiên văn học Carl Sagan và một số người khác thiết kế, thông điệp này chứa thông tin về DNA của con người, dân số và vị trí của Trái đất. Sóng radio này được hướng đến chòm sao M13, cách chúng ta khoảng 25.000 năm ánh sáng. Sóng radio có tốc độ ánh sáng nên thông điệp Arecibo sẽ không đến đó trước 25.000 năm ánh sáng, và thậm chí sau đó nó cũng không đến được. Đó là vì M13 sẽ di chuyển khỏi vị trí của nó năm 1974, theo hệ quy chiếu Trái đất.[14] Tất nhiên là đội nghiên cứu Arecibo biết điều này, nhưng dù sao họ vẫn lợi dụng cơ hội này để quảng bá.

Các chòm sao khác có thể vẫn là những cái đích tốt hơn cho các cuộc thăm dò bằng đài thiên văn vô tuyến. Và trí thông minh phát hiện được có thể sẽ không thuộc về sinh giới.

Sự khẳng định này đến từ SETI (Search for Extra-Terrestrial Intelligence: Tìm kiếm Trí tuệ ngoài hành tinh). Có trụ sở tại Mountain View, California, chỉ cách Google vài dãy phố, tổ chức hiện đã 50 tuổi này đang cố gắng dò tìm các tín hiệu của trí thông minh ngoài hành tinh, phát đi từ các khoảng cách xa đến 100 triệu triệu dặm. Để bắt được sóng radio trong vũ trụ, họ đã đặt 42 đài thiên văn vô tuyến khổng lồ hình đĩa, cách San Francisco 300 dặm về phía bắc. SETI *lắng nghe* các tín hiệu – nó không gửi chúng đi, và trong một nữa thế kỷ qua nó không nghe được gì từ ET (Extra-Terrestrial). Tuy nhiên, họ đã xác lập được một điều chắc chắn

nhưng khó hiểu, khiến ASI có thể bành trướng mà không gặp trở ngại: Ngân hà của chúng ta chỉ có rất ít sinh vật sinh sống, và không ai hiểu tại sao.

Giám đốc thiên văn của SETI, Tiến sĩ Seth Shostak, có một lập trường táo bạo về chuyện chính xác thì chúng ta sẽ tìm thấy *gì*, nếu đến một ngày nào đó ta tìm thấy. Nó sẽ bao gồm trí tuệ nhân tạo, chứ không phải trí tuệ sinh học.

Ông nói với tôi, "Cái chúng ta tìm kiếm ở ngoài kia là một thứ liên tục tiến hóa. Những tiến bộ công nghệ đã dạy chúng ta rằng không có gì bất biến trong thời gian dài. Sóng radio, thứ chúng ta đang lắng nghe, được các thực thể *sinh hoc* tạo ra. Khoảng thời gian từ lúc bạn bắt đầu biết đến sóng radio cho đến khi bạn bắt đầu chế tạo các máy móc tốt hơn chính bạn, thứ biết suy nghĩ, chỉ là vài thế kỷ. Không hơn. Vậy là bạn đã phát minh ra những kẻ thừa kế mình."

Nói cách khác, có một khoảng thời gian tương đối ngắn giữa các cột mốc công nghệ về phát minh ra sóng radio và phát minh ra AI cao cấp của bất cứ dạng sống thông minh nào. Một khi bạn đã phát triển AI cao cấp, nó sẽ chiếm lấy hành tinh hoặc hợp nhất với những người tạo ra sóng radio. Sau đó, radio không cần thiết nữa.

Hầu hết đài thiên văn vô tuyến của SETI đều ngắm đến "vùng Goldilocks"◎ của các ngôi sao gần với Trái đất. Vùng đó đủ gần với sao chủ để các loại chất lỏng trên bề mặt của nó không bị bay hơi hoặc đóng băng. Cần phải "đúng như thế" thì mới có sự sống, vì vậy người ta mới lấy thuật ngữ từ câu chuyện "Goldilocks và ba con gấu."

Shostak lập luận rằng SETI nên hướng *một số* máy thu về phía các góc của Ngân hà, nơi phù hợp với trí tuệ nhân tạo hơn là trí tuệ sinh học trong vũ trụ, một "vùng Goldilocks" của AI. Những khu vực này có mật độ năng lượng đậm đặc — những sao trẻ, sao neutron và lỗ đen.

"Tôi nghĩ chúng ta nên dùng ít nhất là vài phần trăm thời gian để tìm kiếm ở những vùng chắc không hấp dẫn lắm đối với các trí thông minh sinh học, nhưng có thể là chỗ máy móc có ý thức đang cư ngụ. Máy móc có những nhu cầu khác sinh vật. Chúng không có một giới hạn cụ thể cho thời gian tồn tại, và do đó dễ dàng thống trị các trí thông minh khác trong vũ trụ. Vì chúng có thể tiến hóa trong một khoảng thời gian rất, rất ngắn so với các tiến hóa sinh học, nên rất có khả năng là những máy móc thông minh đầu tiên đã hoàn toàn thống trị các trí thông minh trong Ngân hà. Đó là kịch bản 'kẻ thắng làm vua.' "[15]

Shostak đã chỉ ra mối tương quan giữa mạng máy tính đám mây hiện đại, kiểu như của Google, Amazon và Rackspace, với kiểu môi trường siêu lạnh năng lượng cao mà các máy móc siêu thông minh sẽ cần. Chẳng hạn như tinh vân Bok globule – đám mây bụi và khí tối đen có nhiệt độ ở mức -441°F◎, thấp hơn nhiệt độ của không gian vũ trụ gần 200°F.[16] Cũng như những mạng lưới máy tính đám mây của Google ngày nay, các máy móc biết tư duy của tương lai sẽ tỏa nhiều nhiệt và cần được duy trì ở nhiệt độ thấp để tránh nguy cơ nóng chảy.

Khẳng định của Shostak về nơi để tìm ra AI cho ta thấy ý tưởng về việc trí tuệ nhân tạo rời bỏ Trái đất đi tìm các nguồn tài nguyên đã gợi lên những tưởng tượng sâu sắc hơn cả những gì Omohundro và các đồng nghiệp ở MIRI từng nghĩ. Tuy nhiên, không giống họ, Shostak không cho rằng siêu trí tuệ nhân tạo sẽ trở nên nguy hiểm.

"Nếu chúng ta chế tạo được một cỗ máy có trí năng tương đương với một con người, thì trong vòng năm năm những phiên bản đời sau của nó sẽ trở nên thông minh hơn tất cả nhân loại cộng lại. Sau một hoặc hai thế hệ, chúng đơn giản là mặc kệ chúng ta. Cũng như cách bạn mặc kệ những con kiến ở sân sau nhà bạn. Bạn không quét sạch chúng, bạn không biến chúng thành thú nuôi, chúng không có ảnh hưởng gì nhiều đến cuộc sống thường nhật của bạn, nhưng chúng vẫn ở đó."

Vấn đề là, tôi *quét sạch* kiến ở sân sau nhà tôi, đặc biệt khi chúng sắp hàng vào bếp. Sự khác biệt giữa hai lập luận là ở chỗ — ASI sẽ đi vào Ngân hà, hoặc gửi các tàu thăm dò, bởi nó đã dùng hết tài nguyên mà nó cần trên Trái đất, hoặc nó tính được chúng sắp bị dùng hết và cần phải khám phá vũ trụ dù tốn kém. Và nếu như thế thật, thì tại sao chúng ta vẫn tồn tại, trong khi giữ cho chúng ta tồn tại có thể sẽ tiêu tốn một lượng lớn tài nguyên? Và đừng quên là bản thân chúng ta cũng được cấu thành từ các loại vật liệu mà ASI có lẽ muốn dùng vào việc khác.[17]

Tóm lại, để kết cục có hậu của Shostak trở nên khả dĩ, siêu trí tuệ nhân tạo sẽ phải *muốn* chúng ta sống. Nếu chỉ mặc kệ chúng ta thì chưa đủ. Và cho đến nay chưa có một hệ thống đạo đức nào được chấp nhận, cũng như chưa có một phương pháp rõ ràng nào để đưa nó vào kết cấu AI cao cấp.

Nhưng đã có một khoa học non trẻ để hiểu các hành vi của tác nhân siêu trí tuệ nhân tạo. Omohundro là người khởi xướng nó.

• • •

Vậy là chúng ta đã khảo sát ba động lực mà Omohundro cho rằng sẽ thúc đẩy các hệ thống tự nhận thức, tự cải tiến: hiệu suất, tự bảo tồn, và

chiếm đoạt tài nguyên. Chúng ta đã thấy bằng cách nào mà những động lực này lại tạo ra những hậu quả thật tồi tệ khi không có kế hoạch và không được lập trình cực kỳ cẩn thận. Và chúng ta buộc phải tự hỏi: liệu chúng ta có khả năng làm một công việc chính xác như vậy không? Liệu bạn, cũng như tôi, có quan sát thế giới này khi những tai nạn gây thiệt hại khủng khiếp xảy ra, và tự hỏi làm thế nào mà chúng ta có thể làm đúng ngay lần đầu tiên với các AI cực mạnh? Trước những thảm họa hạt nhân ở đảo Three Mile, Chernobyl, Fukushima, chẳng phải các nhà thiết kế và quản trị trình độ cao cũng đã cố gắng hết sức để đề phòng các sự cố đó sao? Thảm họa hạt nhân Chernobyl năm 1986 xảy ra trong một cuộc kiểm tra *an toàn*.[18]

Cả ba thảm họa này đều được nhà lý thuyết về tổ chức Charles Perrow gọi là những "tai nạn thông thường." Trong cuốn *Normal Accidents: Living with High-Risk Technologies* (Tai nạn thông thường: Sống chung với các công nghệ rủi ro cao), Perrow đề xuất rằng các tai nạn, thậm chí thảm họa, là một đặc điểm của những cơ sở hạ tầng phức tạp. Chúng có "tính không thể hiểu được" cao vì chúng chứa đựng những khiếm khuyết trong nhiều quá trình thường không liên quan đến nhau. Các sai lầm riêng rẽ – nếu chỉ từng cái thì sẽ không gây hậu quả nặng nề– sẽ kết hợp lại để tạo ra những hỏng hóc trên cả hệ thống mà không thể đoán trước được.

Tại đảo Three Mile ngày 28/3/1979, bốn sự cố đơn giản tạo nên thảm họa: hai máy bơm của hệ thống làm lạnh dừng hoạt động vì lỗi kỹ thuật; hai máy bơm nước khẩn cấp không thể hoạt động vì van của chúng được đóng để bảo trì; một cái nhãn sửa chữa che mất đèn báo lỗi của hệ thống; một cái van khí lạnh bị kẹt ở trạng thái mở, và một đèn báo bị hỏng lại chỉ ra rằng cái van đó đang đóng. Kết quả tổng hợp: phần lõi của lò phản ứng

bị nấu chảy ra, những thiệt hại về người chỉ được ngăn chặn trong gang tấc, và đó là một cú đánh mạnh vào ngành năng lượng hạt nhân của Hoa Kỳ.

Perrow viết: "Chúng ta đã thiết kế những thứ cực kỳ phức tạp đến nỗi chúng ta không thể tiên liệu hết tất cả tương tác có thể giữa các hỏng hóc chắc chắn sẽ xảy ra; chúng ta thêm vào những thiết bị an toàn mà sẽ bị bỏ qua, bị đánh bại hoặc bị các đường đi ẩn giấu trong hệ thống đánh lừa."[19]

Những hệ thống mà các thành phần của nó "gắn chặt vào nhau," Perrow viết, nghĩa là chúng có những ảnh hưởng ngay lập tức và nghiêm trọng đến nhau, sẽ đặc biệt dễ tổn thương. Một ví dụ nhãn tiền về hiểm họa của các hệ thống AI gắn chặt vào nhau đã xảy ra vào tháng 5/2010 ở phố Wall.

Có đến 70% các giao dịch chứng khoán trên phố Wall được thực hiện nhờ khoảng 80 hệ thống giao dịch tần số cao (high-frequency trading system: HFT) bằng máy tính. Nó tương đương một tỉ cổ phiếu một ngày. Những thuật toán giao dịch và các siêu máy tính đang chạy chúng được các ngân hàng, quỹ đầu cơ và công ty tồn tại chỉ để thực hiện các giao dịch tần số cao sở hữu. Mục đích của HFT là kiếm tiền từ những cơ hội chỉ xảy ra trong nháy mắt. Ví dụ, khi giá của một cổ phiếu thay đổi và giá của những cổ phiếu khác đáng nhẽ phải thay đổi theo ngay nhưng chưa kịp – và mỗi ngày, chớp lấy các cơ hội như vậy nhiều nhất có thể.[20]

Vào tháng 5/2010, Hy Lạp lâm vào khó khăn vì nợ công. Những nước châu Âu đã cho Hy Lạp vay tiền lo ngại nước này tuyên bố vỡ nợ. Cuộc khủng hoảng nợ làm suy yếu nền kinh tế châu Âu, và khiến thị trường Mỹ trở nên mong manh. Tất cả những gì cần thiết để châm ngòi cho một tai họa là sự hoảng sợ của một doanh nhân từ một công ty đầu tư không rõ tên.

Anh ta đặt lệnh bán ngay lập tức 4,1 tỉ đô-la hợp đồng tương lai và chứng chỉ các quỹ ETF (exchange-traded fund) có liên quan đến châu Âu.

Sau lệnh bán đó, giá của các hợp đồng tương lai (E-Mini S&P 500) giảm 4% trong bốn phút. Các giải thuật giao dịch tần số cao (high-frequency-trade algorithm) phát hiện ra sự rớt giá này. Để chốt lời, chúng tự động khởi động việc bán ra, xảy ra chỉ trong vài phần ngàn giây (lệnh bán hoặc mua nhanh nhất hiện tại là 3 mili giây hay 3/1.000 giây). Giá thấp hơn lại tự động khởi động các giao dịch HFT *khác mua vào* E-Mini S&P 500, và bán các tài sản khác để có tiền mua nó. Trước khi con người kịp can thiệp, một phản ứng dây chuyền đã đánh tụt chỉ số Dow-Jones xuống 1.000 điểm.[21] [22] Tất cả xảy ra trong vòng 20 phút.

Perrow gọi vấn đề này là "tính không thể hiểu được." Một tai nạn thông thường bao gồm các tương tác, vốn "không chỉ không mong đợi, mà còn không thể hiểu được trong khoảng thời gian nguy cấp."[23] Không ai đoán trước được các giải thuật sẽ ảnh hưởng lẫn nhau thế nào, không ai hiểu được điều gì đang diễn ra.

Nhà phân tích rủi ro tài chính Steve Ohana thừa nhận vấn nạn này. "Nó là một rủi ro mới xuất hiện," ông nói. "Chúng tôi biết rằng nhiều thuật toán tương tác lẫn nhau, nhưng chúng tôi không biết chính xác theo cách nào. Tôi nghĩ chúng ta đã đi quá xa trong việc điện toán hóa nền tài chính. Chúng ta không thể kiểm soát được con quái vật mà mình đã tạo ra."[24]

Con quái vật đó lại tấn công vào ngày 1/8/2012. Một giải thuật HFT được lập trình tồi đã khiến Công ty đầu tư Knight Capital Partners mất 440 triệu đô-la chỉ trong vòng 30 phút.[25]

Những tai họa này có các yếu tố giống như kiểu của AI mà tôi dự đoán: các hệ thống AI có độ phức tạp cao, hầu như không thể hiểu, các tương tác không đoán trước được với các hệ thống khác và với hệ sinh thái công nghệ thông tin rộng lớn hơn, những lỗi xuất hiện trong tốc độ cực nhanh của máy tính khiến can thiệp của con người trở nên vô nghĩa.

• • •

"Một tác nhân chỉ tìm kiếm sự tăng cường hiệu suất, sự tự bảo tồn và chiếm đoạt tài nguyên thì sẽ hành động như một kẻ tâm thần hoang tưởng và ám ảnh," Omohundro viết trong "Bản chất của trí tuệ nhân tạo tự cải tiến."[26] Có vẻ chỉ làm mà không chơi đã biến AI thành kẻ sống chung nguy hiểm. Một robot chỉ có những động lực như chúng ta đã thảo luận sẽ là một Thành Cát Tư Hãn bằng máy, chiếm đoạt mọi tài nguyên trong Ngân hà, tiêu diệt mọi kẻ cạnh tranh trong đời sống và tàn sát các kẻ thù chưa bộc lộ mối đe dọa cho đến cả ngàn năm sau. Và vẫn còn một động lực nữa để thêm vào cái đống hỗn độn đó – sự sáng tạo.

Động lực thứ tư của AI sẽ khiến hệ thống phát sinh ra những cách mới, hiệu quả để đạt được mục tiêu, hoặc đúng hơn để tránh các kết cục trong đó những mục tiêu của nó không được thỏa mãn một cách tốt nhất. Động lực sáng tạo nghĩa là hệ thống sẽ trở nên khó dự đoán hơn (thật đáng sợ) vì những ý tưởng sáng tạo là những ý tưởng *độc đáo*.[27] Hệ thống càng thông minh thì con đường nó đi càng mới mẻ, và mọi sự lại càng vượt xa tầm với của chúng ta. Động lực sáng tạo sẽ giúp tối đa hóa các động lực khác – hiệu suất, tự bảo tồn, chiếm đoạt tài nguyên – giúp nó đi vòng và vượt lên khi các động lực khác của nó bị cản trở.

Ví dụ, giả sử rằng robot chơi cờ của bạn có mục tiêu chính là đánh thắng bất kỳ đối thủ nào. Khi thi đấu với một robot chơi cờ khác, nó ngay lập tức hack vào CPU của robot kia và giảm tốc độ vi xử lý của con kia xuống tốc độ rùa bò để có một lợi thế tuyệt đối. Bạn phản ứng lại, khoan đã, đây không phải cái tôi muốn. Bạn sửa chữa mã của robot, thêm vào đó một mục tiêu phụ là không được hack CPU của các đối thủ, nhưng trước khi có trận tiếp theo, bạn khám phá ra robot của bạn *đang chế tạo* một robot trợ lý để hack CPU đối thủ hộ nó! Khi bạn cấm nó chế tạo robot, nó liền *thuê* con khác! Nếu không có những hướng dẫn tỉ mỉ, liên tục bổ đắp, thì một hệ thống hướng đích tự nhận thức, tự cải tiến sẽ đi rất xa để đạt các mục tiêu của nó, đến mức chúng ta thấy lố bịch.

Trên đây là một ví dụ về hậu quả khó lường của AI, một vấn đề lớn và rộng đến nỗi nói về nó thì cũng như nói về "thủy nạn" đối với các con tàu đi biển. Một hệ thống AI mạnh có nhiệm vụ bảo đảm an toàn cho bạn nhưng lại có thể giam bạn trong nhà. Nếu bạn muốn được hạnh phúc, nó có thể buộc chặt bạn vào các loại máy trợ năng và kích thích trung khu thần kinh hoan lạc của bạn liên tục. Nếu bạn không cung cấp cho AI một thư viện khổng lồ về các cách hành xử đúng đắn hoặc một nguyên tắc sắt thép để nó hiểu được bạn sẽ thích cách hành xử nào, bạn sẽ phải chấp nhận bất cứ hành vi nào nó chợt nghĩ ra. Và vì nó là một hệ thống cực kỳ phức tạp, nên có thể bạn sẽ không bao giờ hiểu hết nó để chắc chắn rằng bạn đã đúng. Có lẽ phải cần cả một AI mới *khác,* thông minh hơn để xác định liệu robot AI của bạn có định buộc bạn vào giường, cắm các dây điện tử vào tai bạn trong một cố gắng giúp bạn an toàn và hạnh phúc hay không.

Có một cách quan trọng khác để xem xét vấn đề động lực của AI, một cách thích hợp hơn với những người suy nghĩ lạc quan như Omohundro.

Những động lực này biểu thị cho các cơ hội – các cánh cửa mở ra cho nhân loại và khát vọng của họ, chứ không phải đóng sập lại. Nếu chúng ta không muốn hành tinh này và cuối cùng là Ngân hà này bị thống trị bởi những thực thể tuyệt đối vị kỷ và không ngừng nhân bản, với bản tính của Thành Cát Tư Hãn luôn muốn tiêu diệt các dạng sống khác và tiêu diệt lẫn nhau, thì các nhà chế tạo AI phải đặt cho hệ thống của họ những mục tiêu nhân văn. Danh sách ước mơ của Omohundro là: "làm con người hạnh phúc," "viết những bài hát hay," "mua vui cho người khác," "tạo ra thứ toán học sâu sắc," và "sáng tạo nghệ thuật đầy cảm xúc."[28] Rồi lùi lại và quan sát. Với những mục tiêu này, động lực sáng tạo của AI sẽ phát huy tối đa tác dụng và làm giàu đẹp cuộc sống.

"Thế còn nhân loại có đáng để bảo tồn không?" là một câu hỏi quan trọng và hết sức thú vị, câu hỏi mà con người chúng ta đã tự đặt ra dưới nhiều hình thức khác nhau trong một thời gian dài. Cái gì làm nên một cuộc sống tốt? Cái gì là dũng cảm, chính trực, ưu tú? Nghệ thuật nào là hay và âm nhạc nào là đẹp? Sự cần thiết phải làm rõ các giá trị của chúng ta là một trong những cách mà quá trình đi tìm trí tuệ nhân tạo phổ quát sẽ khiến chúng ta hiểu rõ mình hơn. Omohundro tin rằng cuộc khám phá sâu sắc bản chất con người này sẽ cho ra đời những công nghệ phong phú mà không đáng sợ. Anh viết: "Với cả logic lẫn nguồn cảm hứng, chúng ta có thể hướng tới xây dựng một công nghệ nâng đỡ tâm hồn con người thay vì phá hủy nó."[29]

• • •

Tất nhiên là tôi có cái nhìn khác – tôi không chia sẻ sự lạc quan của Omohundro. Nhưng tôi đánh giá cao tầm quan trọng đặc biệt của việc phát

triển một khoa học để hiểu được những máy móc thông minh. Lời cảnh báo của anh về AI cao cấp cần được nhắc lại:

> Tôi không nghĩ rằng phần đông các nhà nghiên cứu AI cho rằng sẽ có bất cứ nguy hiểm gì trong việc chế tạo một robot chơi cờ. Nhưng phân tích của tôi cho thấy rằng chúng ta nên nghĩ thật kỹ về những giá trị chúng ta sẽ đưa vào, nếu không chúng ta sẽ thu được một thực thể bệnh hoạn, ích kỷ và tự coi mình là trung tâm.

Kinh nghiệm của riêng tôi cho thấy anh đã đúng khi nói về những nhà chế tạo AI – những người tôi đã nói chuyện, những người đang tích cực và bận rộn chế tạo các hệ thống thông minh, không nghĩ rằng điều họ đang làm là nguy hiểm. Tuy nhiên, hầu hết đều có một niềm tin sâu sắc rằng trí tuệ nhân tạo sẽ thay thế cho trí tuệ con người.

Nhưng họ lại không suy xét xem điều đó sẽ xảy ra thế nào.

Các nhà chế tạo AI (cũng như các nhà lý thuyết và nhà đạo đức) có xu hướng tin rằng các hệ thống thông minh sẽ chỉ làm những gì mà chúng được lập trình để làm.

Riêng Omohundro thì nói rằng chúng sẽ làm thế, và còn hơn thế nhiều, và chúng ta có thể biết tương đối chính xác việc các hệ thống AI cao cấp sẽ cư xử ra sao. Một số hành vi không được mong đợi và có tính sáng tạo. Điều đó được ẩn trong một khái niệm đơn giản không ngờ, cần có một nhãn quan sâu sắc như của Omohundro để nhìn thấu: *đối với một hệ thống đủ thông minh, né tránh các lỗ hổng là một động lực quan trọng giống như những mục tiêu lớn và nhỏ đã được lập trình sẵn*.

Chúng ta phải đề phòng các hệ quả không lường trước của những mục tiêu mà chúng ta đưa vào các hệ thống thông minh, và cũng phải chú ý đến hệ quả của những thứ mà chúng ta đã bỏ qua.

7

Sự Bùng Nổ Trí Thông Minh

> Theo lập trường v`êhiểm họa diệt vong, một trong những điểm mấu chốt v`êtrí tuệ nhân tạo là một trí tuệ nhân tạo có thể gia tăng trí thông minh của nó cực nhanh. Lý do dễ thấy để nghi ngờ v`êkhả năng này là quá trình tự cải tiến có tính đệ quy. (Good 1965). AI trở nên thông minh hdn, cả trong việc viết các chức năng nhận thức nội tại của AI, cho nên AI sẽ viết lại các chức năng nhận thức đang có để hoạt động tốt hơn. Rồi đi`âu đố lại làm AI thông minh hơn, cả trong việc tự lập trình cho nó, do đó nó lại tạo ra nhi`âu sự cải tiến hơn… Hàm ý chính trong những lập luận này là khi AI đạt đến một cấp độ nhất định, nó có thể thực hiện một bước nhảy vọt khổng l`ôtrong trí thông minh.[1]
>
> — **Eliezer Yudkowsky**
> nhà nghiên cứu, Viện Nghiên cứu Trí thông minh Máy tính

> Có phải ý bạn là: recursion?
>
> — kết quả tìm kiếm trên Google khi gõ "recursion"[◎]

Cho đến nay, trong cuốn sách này, chúng ta đã xem xét một kịch bản v`ề AI, nó tệ đến mức phải nghiên cứu lại thật cẩn thận. Chúng ta đã khảo sát

một ý tưởng đầy hứa hẹn về chuyện xây dựng một AI để tháo gỡ quả bom này – AI thân thiện – và thấy rằng nó chưa hoàn chỉnh. Trên thực tế, nói chung ý tưởng về việc viết mã cho một hệ thống thông minh với những mục tiêu an toàn vĩnh cửu, hoặc có năng lực tạo ra được các mục tiêu an toàn có khả năng sống sót qua một số lớn những vòng lặp tự cải tiến, dường như chỉ là ước muốn viển vông.

Tiếp đến, chúng ta đã khám phá tại sao AI lại nguy hiểm. Chúng ta nhận thấy rằng nhiều động lực của các hệ thống tự nhận thức, tự cải tiến có thể dễ dàng dẫn đến những kết cục tai họa cho con người. Những kết cục đó làm nổi bật mối hiểm họa gần như mang tính nghi thức của tội phạm điều cấm và không làm điều cần thiết trong công cuộc lập trình vốn đầy sai sót của con người.

AGI, một khi đạt được, sẽ nguy hiểm và không thể đoán trước, nhưng trong khoảng thời gian ngắn có lẽ chưa phải là thảm họa. Thậm chí nếu một AGI tự nhân bản và cùng nhau thoát ra, nó sẽ không sở hữu một sự nguy hại tiềm năng hơn so với một nhóm người thông minh. Sự nguy hiểm tiềm ẩn của AGI nằm trong phần cốt lõi của kịch bản Đứa trẻ Bận rộn, khi quá trình tự cải tiến đệ quy tốc độ cao khiến AI từ một trí tuệ nhân tạo phổ quát tự vươn lên thành siêu trí tuệ nhân tạo. Nó thường được gọi là "sự bùng nổ trí thông minh."

Một hệ thống tự nhận thức, tự cải tiến sẽ tìm cách để đạt được trọn vẹn các mục tiêu của nó và giảm thiểu các lỗ hổng bằng cách tự nâng cấp. Nó sẽ tìm không chỉ các cải tiến nhỏ, mà cả các cải tiến lớn, liên tục trên mọi khía cạnh của những kỹ năng nhận thức, đặc biệt là những kỹ năng liên quan và tạo ra tiến bộ mới trong trí thông minh của nó. Nó sẽ tìm kiếm một trí thông minh tốt hơn con người, hay còn gọi là siêu thông minh. Nếu

không có những cách lập trình khéo léo, chúng ta sẽ phải đối diện với nguy cơ đến từ các cỗ máy siêu thông minh.

Theo Steve Omohundro, chúng ta biết rằng AGI sẽ tìm kiếm sự bùng nổ trí thông minh một cách tự nhiên. Nhưng sự bùng nổ trí thông minh chính xác là gì? Những yêu cầu tối thiểu về phần cứng và phần mềm ra sao? Liệu các yếu tố như thiếu hụt nguồn vốn hoặc sự phức tạp của việc đạt được trí thông minh máy tính có chặn đứng quá trình bùng nổ trí thông minh từ trong trứng nước?

Trước khi khảo sát cơ chế của sự bùng nổ trí thông minh, sẽ là quan trọng để biết chính xác thuật ngữ đó nghĩa là gì, và làm thế nào mà ý tưởng về sự bùng nổ trí tuệ nhân tạo lại được nhà toán học I. J. Good đề xuất và phát triển.

Con đường quốc lộ số 81 (I-81) bắt đầu từ bang New York và kết thúc ở Tennessee, vắt ngang qua hầu hết rặng núi Appalachia. Từ vùng giữa bang Virginia kéo xuống phía nam, con đường cao tốc này uốn lượn giữa những dải rừng rậm đồi núi và những đồng cỏ xanh mướt lộng gió, xuyên qua các khung cảnh nguyên sơ lúc ấn tượng lúc mơ hồ của nước Mỹ. Trong rặng Appalachia có dãy Blue Ridge (từ bang Pennsylvania đến Georgia) và dãy Great Smokies (dọc theo biên giới giữa bang Bắc Carolina và Tennessee). Càng đi về phía nam, bạn càng khó bắt sóng điện thoại, số nhà thờ dường như còn nhiều hơn nhà ở, và âm nhạc trong radio chuyển từ nhạc đồng quê sang nhạc phúc âm, rồi sang các bài thuyết giảng về hỏa ngục. Tôi nghe một bài hát khó quên về sự cám dỗ tên là "Long Black Train" (Con tàu dài và đen) của Josh Turner. Tôi nghe một giáo sĩ bắt đầu bài giảng đạo về Abraham và Isaac, chẳng hiểu ông ấy nói gì, và kết thúc với truyện ngụ ngôn về những ổ bánh mì, cá và *địa ngục*, được thêm vào

để tạo hiệu ứng. Tôi đến gần dãy Great Smokies, biên giới Bắc Carolina, ở đó có Đại học Virginia Tech (Virginia Polytechnic Insitute and State University) ở Blacksburg, Virginia. Khẩu hiệu của trường là Invent the Future (Sáng tạo tương lai).

20 năm trước, nếu lái xe trên con đường I-81 gần giống như ngày nay, xe bạn có thể bị một chiếc xe mui trần hiệu Triumph Spitfire mang biển số 007 IJG vượt qua. Cái biển số xe cao ngạo này thuộc về I. J. Good, đến Blacksburg vào năm 1967 với tư cách một giáo sư đầu ngành về thống kê. Con số "007" là để liên tưởng tới Ian Fleming© và nhiệm vụ bí mật của Good trong Thế chiến II với tư cách một người giải mật mã ở Bletchley Park, Anh. Việc phá được hệ thống mật mã mà quân đội Đức Quốc xã dùng để truyền tin đã góp phần không nhỏ vào thất bại của phe Trục. Tại Bletchley Park, Good làm việc cùng với Alan Turing, được coi là cha đẻ của ngành điện toán hiện đại (và cũng là người tạo ra bài kiểm tra Turing mà chúng ta đã đề cập ở Chương 4), ông đã chế tạo, lập trình một trong những chiếc máy tính điện tử đầu tiên.

Tại Blacksburg, Good là một giáo sư nổi tiếng – lương của ông còn cao hơn hiệu trưởng của trường đại học. Là một người cuồng số học, ông nhận thấy mình đã đến Blacksburg vào giờ thứ 7 của ngày thứ 7 của tháng thứ 7 của năm thứ 7 của thập niên thứ 7, và được phân cho căn hộ số 7 của nhà số 7 của dãy nhà Terrace View. Good nói với bạn bè mình rằng Chúa đã quẳng những sự trùng hợp đó vào một kẻ vô thần như ông để cảnh tỉnh ông về sự hiện hữu của người.

"Tôi có một ý tưởng mơ hồ rằng anh càng nghi ngờ sự hiện hữu của Chúa, người sẽ càng cung cấp nhiều sự trùng hợp, bằng cách đó đưa cho

anh bằng chứng về đức tin mà không phải ép buộc anh," Good nói. "Khi tôi tin vào lý thuyết đó, những sự trùng hợp có lẽ sẽ ngừng lại."[2]

Tôi đã đến Blacksburg để tìm hiểu về Good, người mới mất gần đây ở tuổi 92, theo lời kể của bạn bè ông. Tôi chủ yếu muốn biết bằng cách nào mà I. J. Good đã nghĩ ra ý tưởng về sự bùng nổ trí thông minh, và liệu nó có khả năng xảy ra không. Sự bùng nổ trí thông minh là mắt xích đầu tiên trong dây chuyền ý tưởng đã làm nên giả thuyết về Singularity.

Không may là trong một tương lai gần, nói đến Virginia Tech cũng có nghĩa là đề cập đến cuộc thảm sát Virginia Tech. Tại đây vào ngày 16/4/2007, Seung-Hui Cho, một sinh viên Anh ngữ đã giết 32 sinh viên và giáo viên và làm bị thương 25 người khác. Đấy là vụ tàn sát bằng súng dã man nhất do chỉ một người thực hiện trong lịch sử Mỹ. Những chi tiết chính của vụ việc là Cho đã bắn chết một nữ sinh tại giảng đường Ambler Johnston thuộc khuôn viên Virginia Tech, rồi giết một nam sinh đang chạy đến giúp. Hai giờ sau, Cho bắt đầu điên cuồng bắn giết những người khác. Trừ hai người đầu tiên, y đã bắn các nạn nhân tại giảng đường Norris của Đại học Virginia Tech. Trước khi bắt đầu bắn, Cho lấy xích khóa trái những cánh cửa gỗ sồi dày của giảng đường để không cho ai trốn thoát.

Khi người bạn lâu năm và là đồng nghiệp ngành thống kê của I. J. Good – Tiến sĩ Golde Holtzman dẫn tôi đi xem văn phòng trước đây của ông ở giảng đường Hutcheson, nằm ở phía bên kia khu vườn tươi đẹp Drillfield (trước vốn là nơi diễu binh), tôi nhận thấy bạn có thể dễ dàng nhìn qua cửa sổ phòng ông sang giảng đường Norris. Nhưng vào thời điểm thảm kịch diễn ra, Holtzman nói với tôi, Good đã nghỉ hưu. Ông không ở trong văn phòng của mình mà đang ở nhà, có lẽ đang tính toán khả năng tồn tại của Chúa.

Theo Tiến sĩ Holtzman, trước khi mất ít lâu. Good đã nâng khả năng đó từ 0 lên 0,1. Ông làm thế, vì với tư cách một nhà thống kê, ông cũng là một người theo trường phái Bayes lâu năm. Được đặt tên theo Thomas Bayes, nhà thống kê học kiêm mục sư thế kỷ 18, ý tưởng chính của thống kê học Bayes là khi tính toán khả năng của một phát biểu nào đó, bạn có thể bắt đầu với niềm tin cá nhân. Sau đó bạn sẽ cập nhật niềm tin đó khi các bằng chứng mới được tìm ra, chúng hỗ trợ hoặc chống lại phát biểu đó.

Nếu Good *không tin vào* Chúa 100% thì không dữ liệu nào, kể cả sự xuất hiện của Chúa, có thể thay đổi được ý nghĩ của ông. Do đó, để nhất quán với góc nhìn theo thuyết Bayes của ông, Good đã gán một xác suất dương nhỏ cho sự tồn tại của Chúa để chắc chắn rằng ông có thể học hỏi từ những dữ liệu mới nếu nó xuất hiện.

Trong bài báo *Speculations Concerning the first Ultraintelligent Machine* (Những suy đoán về Cỗ máy siêu thông minh đầu tiên) năm 1965, Good đã trình bày một luận chứng đơn giản và dễ hiểu, luôn được nhắc đến trong các cuộc thảo luận về trí tuệ nhân tạo và Singularity:

> Hãy định nghĩa một cỗ máy siêu thông minh là một cỗ máy có khả năng làm những công việc trí óc giỏi hơn bất cứ người nào, dù anh ta có thông minh đến đâu. Vì việc thiết kế máy móc cũng là một trong những công việc đòi hỏi trí óc, một cỗ máy siêu thông minh sẽ thiết kế những cỗ máy thậm chí còn thông minh hơn; và sau đó, không nghi ngờ gì, sẽ là một "sự bùng nổ trí thông minh," và trí tuệ con người sẽ bị bỏ lại sau rất xa. Do vậy, cỗ máy siêu thông minh đầu tiên sẽ là thứ cuối cùng mà con người cần phát minh ra…[3]

Singularity có ba định nghĩa hoàn chỉnh — định nghĩa trên đây của Good là cái đầu tiên.[4] Good chưa bao giờ sử dụng thuật ngữ "Singularity," nhưng ông đã khởi đầu nó bằng cách khẳng định điều mà ông nghĩ là một cột mốc có lợi và không thể tránh khỏi trong lịch sử loài người – phát minh ra những máy móc thông minh hơn con người. Diễn giải lại ý của Good, nếu bạn tạo ra một cỗ máy siêu thông minh, nó sẽ giỏi hơn con người trong mọi việc mà chúng ta cần dùng đến bộ não của mình, bao gồm cả chuyện chế tạo những cỗ máy siêu thông minh. Cỗ máy đầu tiên sẽ châm ngòi cho sự bùng nổ trí thông minh, một sự gia tăng trí thông minh siêu tốc khi nó liên tục tự cải tiến, hoặc đơn giản là tạo ra các máy móc thông minh hơn. Cỗ máy, hoặc những cỗ máy này sẽ vượt xa sức mạnh bộ não người. Sau sự bùng nổ trí thông minh, nhân loại sẽ không cần phải phát minh ra bất cứ thứ gì khác – tất cả những gì họ cần, máy móc sẽ làm hộ.

Đoạn văn này của Good đã được trích dẫn một cách thích đáng trong các cuốn sách, bài báo và bài luận về Singularity, về tương lai và mối nguy hiểm của trí tuệ nhân tạo. Nhưng hai ý tưởng quan trọng hầu như luôn-bị-bỏ qua. Thứ nhất là câu dẫn nhập đặc biệt của bài. Nó thế này: "Sự tồn tại của con người phụ thuộc vào những bước đầu tiên trong việc xây dựng một cỗ máy siêu thông minh." Thứ hai là *nửa sau* của câu cuối trong đoạn văn trên, vốn rất hay bị bỏ qua. Câu cuối của đoạn văn hay được trích dẫn của Good phải đọc đầy đủ:

> Do vậy, cỗ máy siêu thông minh đầu tiên sẽ là thứ cuối cùng mà cọn người cần phát minh ra, *với giả thiết rằng cỗ máy đó sẽ đủ ngoan ngoãn để nói cho chúng ta biết làm sao để giữ được nó trong vòng kiểm soát*. (in nghiêng là của tôi).

Hai câu trên cho ta biết những điểm quan trọng trong ý kiến của Good. Ông cảm thấy con người chúng ta bị quá nhiều vấn đề tăm tối và phức tạp bao vây – cuộc chạy đua vũ khí hạt nhân, ô nhiễm, chiến tranh… – đến nỗi chúng ta chỉ có thể được cứu nhờ những tư duy siêu việt hơn sẽ đến từ những máy móc siêu thông minh. Câu thứ hai làm ta hiểu cha đẻ của khái niệm sự bùng nổ trí thông minh đã nhận thức một cách sâu sắc rằng việc chế tạo các máy móc siêu thông minh, dù có cần thiết đến đâu đối với sự sống còn của loài người, cũng có thể quay trở lại làm hại chúng ta. Giữ một cỗ máy siêu thông minh trong tầm kiểm soát không phải là chuyện dễ, Good nói với chúng ta như thế. Ông thậm chí không tin rằng chúng ta sẽ biết cách thực hiện điều này – cỗ máy đó sẽ phải *tự nói* cho ta biết.

Good biết một vài điều về những máy móc cứu thế giới – ông đã góp phần xây dựng và vận hành những máy tính điện tử đầu tiên, được sử dụng tại Bietchley Park hòng đánh bại Đức. Ông cũng biết một số điều về hiểm họa diệt vong – ông là một người Do Thái chiến đấu chống lại Đức Quốc xã, bố ông đã tránh được cuộc tàn sát người Do Thái ở Ba Lan bằng cách nhập cư vào Liên hiệp Anh.

Khi còn là một đứa trẻ, bố của Good, một người Ba Lan và là trí thức tự học, đã học nghề đồng hồ bằng cách nhìn trộm các thợ đồng hồ khác qua ô cửa. Ông mới 17 tuổi khi đến Anh vào năm 1903 với 35 rúp trong túi và một chiếc bánh pho mát lớn. Ở London, ông làm những công việc lặt vặt cho đến khi có thể mở cửa hàng trang sức riêng của mình. Ông phát đạt và kết hôn. Năm 1915, Isidore Jacob Gudak (sau đó đổi tên thành Irving John "Jack" Good) chào đời.[5] Tiếp theo là một em trai và một em gái, một vũ nữ tài năng đã mất trong vụ cháy rạp hát. Đau đớn trước cái chết của em gái, Jack Good đã chối bỏ sự tồn tại của Chúa.

Good là một thần đồng toán học, đã có lần đứng trong cũi và hỏi mẹ 1.000 lần của 1.000 bằng bao nhiêu. Trong một lần vật lộn với bệnh bạch hầu, ông đã tự mình khám phá ra số vô tỉ (những số không thể biểu diễn bằng phân số, ví dụ như $\sqrt{2}$). Trước năm 14 tuổi, ông đã tìm ra phép quy nạp, một phương pháp chứng minh toán học. Kể từ đó các giáo viên toán của ông chỉ để ông tự học với những cuốn sách. Tại Đại học Cambridge, Good đã đoạt mọi giải thưởng về toán trong quá trình học tiến sĩ, và tìm thấy niềm đam mê với môn cờ vua.

Do nhận thấy tài năng chơi cờ của ông, một năm sau khi Thế chiến II nổ ra, đương kim vô địch cờ vua của Vương quốc Anh, Hugh Alexander, đã tuyển Good vào Lều số 18 ở Bietchley Park. Lều 18 là nơi các nhà giải mã làm việc. Họ phá các loại mật mã được phe Trục – Đức, Nhật và Ý – sử dụng để truyền đi các mệnh lệnh quân sự, nhưng họ tập trung nhiều nhất vào Đức. Tàu ngầm U-boat của Đức đang đánh chìm tàu của phe Đồng minh với một tốc độ đáng sợ – chỉ trong nửa đầu năm 1942, U-boat đã đánh đắm khoảng 500 tàu Đồng minh.[6] Thủ tướng Anh Winston Churchill lo ngại rằng đảo quốc của ông sẽ bị cô lập đến mức phải đầu hàng.

Những thông điệp của nước Đức được gửi qua sóng radio, và người Anh dò tìm chúng bằng các tháp thu tín hiệu. Ngay từ đầu cuộc chiến, Đức đã mã hóa những thông điệp đó bằng một cỗ máy có tên Enigma. Được phân phối rộng rãi trong lực lượng quân đội Đức, Enigma có kích cỡ và hình dáng giống như một máy chữ thủ công đời cũ. Mỗi phím biểu thị một chữ cái, và được kết nối với một dây kim loại. Sợi dây này sẽ kết nối với một sợi dây khác được gắn với một chữ cái khác. Chữ cái đó sẽ thay thế cho chữ cái đầu tiên để thành chìa khóa. Tất cả các sợi dây được gắn vào những rô-to được thiết kế để một chữ cái có thể kết nối với bất kỳ một chữ

cái nào khác trong bảng chữ cái. Máy Enigma cơ bản có ba đĩa quay, và mỗi đĩa sẽ hoán đổi những ký tự đã được những đĩa trước nó hoán đổi. Đối với bảng chữ cái 26 ký tự, có tất cả 403.291.461.126.605.635. 584.000.000 phép hoán đổi.[7] Những đĩa quay, hay những tổ hợp sắp đặt của máy, được thay đổi gần như hằng ngày.

Khi một người Đức gửi những người khác một thông điệp được Enigma mã hóa, người nhận sẽ dùng máy Enigma của họ để giải mã nó, nếu họ biết tổ hợp sắp đặt của người gửi.

May mắn là Bietchley Park cũng có vũ khí riêng của họ – Alan Turing. Trước chiến tranh, Turing đã học toán và mã hóa tại Đại học Cambridge và Princeton. Ông đã tưởng tượng ra một "cỗ máy tự động," thứ hiện được gọi là máy Turing. Cỗ máy tự động này đã đặt nền móng cho các nguyên tắc đầu tiên của ngành điện toán.

Giả thuyết Church-Turing, tổng hợp các công trình của Turing và giáo sư của ông tại Princeton, nhà toán học Alonso Church, đã thực sự gieo mầm cho khu vườn của ngành trí tuệ nhân tạo. Nó đề xuất rằng bất cứ thứ gì mà một giải thuật, hoặc một chương trình có thể tính toán được, thì một máy Turing cũng có thể tính toán được. Do đó, nếu các quá trình của bộ não có thể biểu thị như một chuỗi tính toán – một giải thuật – thì một máy tính cũng có thể xử lý thông tin y như thế. Nói cách khác, trừ phi có gì đó huyền bí hoặc kỳ diệu trong tư duy của con người, còn không máy tính có thể đạt được tới trí thông minh. Nhiều nhà nghiên cứu AGI đã đặt hy vọng vào giả thuyết Church-Turing.

Chiến tranh đã đem đến cho Turing một khóa học cấp tốc về mọi thứ ông từng nghĩ tới trước cuộc chiến, và nhiều thứ mà ông *chưa từng* quan

tâm đến, như Đức Quốc xã và tàu ngầm. Vào cao điểm chiến tranh, người của Bletchley Park đã giải mã khoảng 4.000 thông điệp mỗi ngày. Phá khóa tất cả chúng bằng tay trở nên bất khả thi. Đấy là một công việc dành cho máy tính. Và Turing đã có một nhận thức quan trọng rằng sẽ dễ hơn nếu ta tìm ra những tổ hợp *không đúng,* thay vì đi tìm những tổ hợp *đúng* trên Enigma.

Những nhà giải mã có dữ liệu để tham khảo – những thông điệp dò được và đã được "phá mã" bằng tay, hoặc bằng các máy phá mã điện tử tên là Bombes. Họ gọi các thông điệp đó là "những nụ hôn." Giống như I. J. Good, Turing là một người sùng bái thuyết Bayes, trong một thời kỳ mà các phương pháp thống kê còn được xem như một phép phù thủy. Cốt lõi của phương pháp này, lý thuyết Bayes, mô tả cách sử dụng dữ liệu để suy ra xác suất của một sự kiện không biết, trong trường hợp này là các tổ hợp của Enigma. Những "nụ hôn" là dữ liệu cho phép các nhà giải mã quyết định các tổ hợp nào sẽ có xác suất không đúng cao, để những nỗ lực giải mã có thể tập trung tốt hơn vào các mảng khác. Tất nhiên, các tổ hợp được thay đổi hằng ngày, nên công việc tại Bletchley Park là một cuộc chạy đua không ngừng nghỉ.

Turing và đồng nghiệp thiết kế một chuỗi các máy điện tử dùng để đánh giá và loại bỏ các tổ hợp có thể của Enigma.[8] Những máy tính đầu tiên này đạt đến trình độ cao nhất trong một loạt cỗ máy cùng tên "Người khổng lồ." Người khổng lồ có thể đọc 5.000 ký tự trong một giây từ một băng giấy chạy qua nó với tốc độ 43 km/h. Nó chứa 1.500 ống chân không, choán đầy một cán phòng. Một trong những người sử dụng chính, đóng góp phân nửa lý thuyết đằng sau Người khổng lồ là nhà thống kê chính của Turing trong hầu hết cuộc chiến: Irving John Good.

Những người hùng tại Bletchley Park có lẽ đã rút ngắn được Thế chiến II từ hai đến bốn năm, cứu sống vô số người.[9] Nhưng không có cuộc diễu hành nào cho những chiến binh thầm lặng ấy. Churchill ra lệnh đập nát tất cả các máy giải mã tại Bletchley thành những cục không lớn hơn một nắm tay, để sức mạnh vô song của chúng không bị lợi dụng chống lại Anh. Những nhà giải mã phải thề giữ kín mọi chuyện trong *30 năm*. Turing và Good được tuyển vào làm việc tại Đại học Manchester, nơi mà lãnh đạo trước đây của họ, Max Newman, có ý định phát triển một máy tính cho mục đích thông thường. Turing đang thiết kế một máy tính tại Phòng nghiên cứu Vật lý Quốc gia thì cuộc sống của ông bị xáo trộn một cách đột ngột. Một người đàn ông có quan hệ đồng tính với ông đã đột nhập vào nhà ông và ăn cắp. Khi thông báo vụ việc này với cảnh sát thì ông cũng thừa nhận mình là người đồng tính. Ông bị buộc tội hủ hóa và bị tước bỏ quyền truy cập vào các tài liệu mật.

Tại Bletchley Park, Turing và Good đã thảo luận những ý tưởng về tương lai như máy tính, máy móc thông minh và máy chơi cờ "tự động." Turing và Good trở nên thân thiết qua những ván cờ vua mà người thắng là Good. Đổi lại, Turing dạy ông chơi cờ vây, một trò chơi chiến thuật của châu Á, và người thắng cũng là Good.[10] Là một vận động viên chạy việt dã tầm cỡ quốc tế, Turing đã lập ra một kiểu chơi cờ mới khiến hai người ngang sức hơn. Cứ sau mỗi nước đi người chơi phải chạy quanh vườn hai vòng. Nếu sau hai vòng về lại bàn cờ mà đối phương vẫn chưa đi nước của mình thì người chạy sẽ được đi *hai nước* liên tiếp.

Bản án về tội hủ hóa của Turing vào năm 1952 làm Good ngạc nhiên, ông không biết Turing là người đồng tính. Turing bị buộc phải chọn giữa nhà tù hoặc tiêm hormone. Ông chọn cái sau, và bị tiêm định kỳ estrogen◎.

Năm 1954, ông ăn một quả táo nhúng trong chất độc cyanide. Có lời đồn vô căn cứ nhưng khá thú vị rằng Apple đã nghĩ ra logo của mình từ bi kịch này.

Sau khi kỳ hạn giữ bí mật chấm dứt, Good là một trong những người đầu tiên lên tiếng chỉ trích cách đối xử của chính phủ đối với người hùng chiến tranh và cũng là bạn hữu của ông.

"Tôi sẽ không nói rằng điều Turing đã làm khiến chúng ta giành chiến thắng," Good nói. "Nhưng tôi dám chắc rằng chúng ta có lẽ đã thất bại nếu không có anh ấy."[11] Năm 1967, Good từ bỏ vị trí tại Đại học Oxford để chuyển sang làm việc tại Đại học Virginia Tech ở Blacksburg, Virginia. Lúc đó ông 52 tuổi. Trong quãng đời còn lại của mình, ông chỉ quay lại Anh đúng một lần.

Đi cùng ông trong chuyến đi năm 1983 là một trợ lý 25 tuổi, đẹp và cao, một cô gái Tennessee tóc vàng tên là Leslie Pendleton. Good gặp Pendleton năm 1980 sau khi ông đã đổi đến 10 thư ký trong 13 năm. Tốt nghiệp tại chính trường Virginia Tech, Pendleton đã không khuất phục trước tính cẩu toàn khó chịu của Good, trụ lại được ở nơi mà những người khác từ bỏ. Sau này, bà kể với tôi về lần đầu tiên đi gửi một trong những bài báo của ông cho một tạp chí toán học, "Ông ấy giám sát cách tôi bỏ tài liệu và thư vào phong bì. Ông ấy giám sát cách tôi dán phong bì, ông ấy không thích dùng nước bọt và bắt tôi dùng một miếng xốp thấm nước. Ông ấy nhìn tôi dán tem. Ông ấy ở đó chờ tôi trở lại từ bưu điện để chắc chắn rằng việc gửi thư đã xong xuôi, cứ như tôi có thể bị bắt cóc hoặc sao đó. Ông ấy là một người đàn ông nhỏ thó kỳ quặc."

Good từng muốn cưới Pendleton. Tuy nhiên, ngay bước đầu tiên là sự khác biệt tuổi tác đến 40 năm, bà đã thấy chẳng thể vượt qua. Song giữa người đàn ông Anh lập dị và người đẹp Tennesse đã nảy sinh một mối gắn kết mà đến tận bây giờ bà vẫn thấy khó mô tả. Trong 30 năm, bà đi cùng ông trong các kỳ nghỉ, quản lý mọi giấy tờ tài liệu của ông, và trợ giúp công việc cho đến khi ông nghỉ hưu, rồi già yếu. Khi chúng tôi gặp nhau, bà đưa tôi đi thăm ngôi nhà của ông ở Blacksburg, một ngôi nhà gạch đơn độc nhìn ra tuyến đường U.S. 460, vốn chỉ là một con đường đồng quê hai làn xe khi Good chuyển tới đây.

Leslie Pendleton đẹp như một pho tượng, giờ đây khoảng 55 tuổi, là tiến sĩ và mẹ của hai đứa con đã lớn. Bà là giáo sư và người quản lý của Virginia Tech, phụ trách các thời khóa biểu, lớp học, và châm biếm các giáo sư, một việc mà bà đã được luyện tập nhiều năm. Ngay cả khi đã cưới một người tầm tuổi bà và có gia đình riêng, nhưng nhiều người trong cộng đồng nơi đây vẫn thấy khó hiểu về mối quan hệ của bà với Good. Họ cuối cùng đã nhận được câu trả lời vào năm 2009 tại đám tang của ông, qua bài điếu văn của Pendleton. Không, họ chưa bao giờ là những người tình, bà nói, nhưng đúng là họ đã dành cả cuộc đời cho nhau. Good không có được một mối tình với Pendleton, nhưng ông đã tìm thấy một người tri kỷ trong suốt 30 năm, một người giám hộ tận tụy đối với những di sản và ký ức về ông.

Đứng trong sân nhà Good, nghe tiếng côn trùng vo ve trên đại lộ 460, tôi hỏi Pendleton, liệu nhà giải mã có bao giờ thảo luận về sự bùng nổ trí thông minh, và liệu máy tính có thể cứu thế giới lần nữa không, như nó đã làm khi ông còn trẻ. Bà suy nghĩ một lát, cố gắng nhớ lại một ký ức đã xa. Sau đó bà trả lời, khá bất ngờ, rằng Good đã thay đổi ý kiến của mình về

sự bùng nổ trí thông minh. Bà phải xem lại những tài liệu của ông trước khi có thể kể tiếp cho tôi.

Tối hôm đó, tại một tiệm Outback Steakhouse© nơi Good và bạn ông là Golde Holtzman thường đứng với nhau vào tối thứ Bảy, Holtzman kể với tôi rằng có ba thứ làm Good trăn trở – Thế chiến II, cuộc thảm sát người Do Thái, và số phận cay đắng của Turing. Điều này khiến tâm trí tôi kết nối giữa công việc của Good trong chiến tranh với những gì ông viết trong *Speculations Concerning the First Ultraintelligent Machine*. Good và đồng nghiệp của ông từng đối đầu với hiểm họa sống còn, và họ đã góp phần đánh bại nó nhờ các máy móc điện toán. Nếu một cỗ máy có thể cứu thế giới vào những năm 1940, thì có lẽ một cỗ máy siêu thông minh có thể giải quyết những vấn đề của nhân loại vào những năm 1960. Và nếu máy móc có thể *học,* trí thông minh của nó sẽ bùng nổ. Nhân loại sẽ phải thích nghi với việc chia sẻ hành tinh này với các máy móc siêu thông minh. Trong *Speculations* ông viết:

> Máy móc sẽ tạo ra các vấn đề xã hội, nhưng chúng cũng có khả năng giải quyết những vấn đề đó, cộng thêm những vấn đề do các loại vi khuẩn và con người tạo ra. Máy móc ấy sẽ khiến người ta sợ hãi và kính nể, có lẽ thậm chí yêu quý. Nhận định này có vẻ hoang đường với một số người đọc, nhưng với người viết, chúng rất thật và cấp bách, đáng để nhấn mạnh rằng đây không phải chuyện khoa học viễn tưởng.

Không có sợi dây liên hệ trực tiếp nào về mặt khái niệm giữa Bletchley Park và sự bùng nổ trí thông minh, nhưng gián tiếp thì có, với nhiều tác động. Trong một cuộc phỏng vấn năm 1996 với nhà thống kê học và là học trò cũ David L. Banks, Good tiết lộ rằng ông đã muốn viết tiểu luận này

sau khi đào sâu nghiên cứu về các mạng neuron nhân tạo. Còn được gọi là các ANN, chúng là những mô hình điện toán mô phỏng hoạt động của các mạng neuron ở con người. Khi bị kích thích, một neuron trong não sẽ bắn tín hiệu đến các neuron khác. Tín hiệu đó sẽ mã hóa một ký ức hay kích hoạt một hành động, hoặc cả hai. Good đã đọc một cuốn sách do nhà tâm lý học Donald Hebb viết năm 1949, trong đó đề xuất rằng hành vi của neuron có thể được mô phỏng bằng toán học.

Một "neuron" điện toán sẽ được kết nối với các neuron điện toán khác. Mỗi kết nối sẽ có một chỉ số "sức nặng," tùy theo độ mạnh của nó. Quá trình máy học sẽ xuất hiện khi hai neuron được kích hoạt cùng lúc, tăng cường "sức nặng" của kết nối giữa chúng. "Các tế bào cùng phát tín hiệu, được kết nối cùng nhau" trở thành khẩu hiệu cho lý thuyết của Hebb. Năm 1957, nhà tâm lý học Frank Rosenblatt của Học viện Công nghệ Massachusetts (MIT) đã tạo ra một mạng neuron dựa theo công trình của Hebb, được ông gọi là "Perceptron." Được xây dựng trên một máy tính IBM có kích cỡ một căn phòng, Perceptron "nhìn thấy" và học các mẫu hình ảnh đơn giản. Năm I960, IBM yêu cầu I. J. Good đánh giá máy Perceptron.[12] "Tôi nghĩ rằng mạng neuron, với cách hoạt động siêu song song của nó, sẽ có thể dẫn đến các máy móc thông minh như con đường lập trình," Good nói.[13] Những cuộc nói chuyện đầu tiên, là cơ sở để Good phát triển thành bài luận *Speculations Concerning the First Ultraintelligent Machine,* đã được trình bày sau đó hai năm.◎ Khái niệm sự bùng nổ trí thông minh ra đời.

Good đã nhìn xa hơn cả những gì ông biết về ANN. Ngày nay, các mạng neuron nhân tạo là ứng viên nặng ký của trí tuệ nhân tạo, được ứng dụng từ công nghệ nhận dạng giọng nói và chữ viết tay cho đến việc xây

dựng mô hình tài chính, phê chuẩn tín dụng và điều khiển robot. ANN hoạt động cực tốt trong những việc đòi hỏi kỹ năng nhận dạng nhanh chóng các mẫu ở trình độ cao. Hầu hết đều bao gồm việc "huấn luyện" mạng neuron bằng một lượng lớn dữ liệu (còn gọi là bộ dữ liệu tập huấn) để mạng này có thể "học" các mẫu. Sau đó nó có thể nhận ra các mẫu tương tự trong dữ liệu mới. Những nhà phân tích có thể hỏi, dựa theo dữ liệu của tháng trước, thị trường chứng khoán sẽ thế nào trong *tuần tới?* Hoặc, khả năng để một người không trả được nợ trên khoản thế chấp ra sao, dựa theo lịch sử về dữ liệu thu nhập, tiêu dùng, tín dụng trong ba năm của người đó?

Giống như các giải thuật di truyền, ANN là các hệ thống "hộp đen." Tức là các dữ liệu đầu vào – sức nặng của các kết nối và sự kích hoạt các neuron – có tính minh bạch. Và các dữ liệu đầu ra cũng hợp lý. Nhưng điều gì đã xảy ra trong đoạn giữa? Không ai biết cả. Các dữ liệu đầu ra của công cụ trí tuệ nhân tạo dạng "hộp đen" không bao giờ có thể tiên đoán được, nên chúng không bao giờ thực sự "an toàn."

• • •

Nhưng nhiều khả năng chúng sẽ đóng một vai trò lớn trong các hệ thống AGI. Ngày nay, nhiều nhà nghiên cứu tin rằng nhận dạng mẫu — thứ mà máy Perceptron của Rosenblatt hướng tới – là công cụ chính làm nên trí thông minh của bộ não chúng ta. Jeff Hawkins, người phát minh ra hai loại thiết bị hỗ trợ cầm tay Palm Pilot và Handspring Treo, đã tiên phong sử dụng ANN trong nhận dạng chữ viết tay. Công ty Numenta của ông có mục tiêu đạt đến AGI bằng công nghệ nhận dạng mẫu. Dileep George, cựu Giám đốc công nghệ của Numenta, giờ là chủ của Vicarious Systems, một công

ty với tham vọng được chỉ rõ trên khẩu hiệu: Chúng tôi đang xây dựng phần mềm biết nghĩ và học như con người.

Nhà khoa học thần kinh, nhà khoa học nhận thức, kiêm kỹ sư sinh dược Steven Grossberg đã đi tới một mô hình dựa trên ANN, được một số người trong ngành tin rằng nó thực sự có thể dẫn tới AGI, và có lẽ tới thứ "cực thông minh" mà tiềm năng của nó đã được Good nhận ra trong các mạng neuron. Nói rộng ra, trước tiên Grossberg xác định vai trò nhận thức của các vùng khác nhau trên vỏ não. Đó là nơi thông tin được xử lý, và các ý nghĩ khởi phát. Sau đó ông tạo ra các ANN mô phỏng từng vùng. Ông đã có những thành tựu trong việc xử lý chuyển động và nói chuyện, phát hiện định dạng và những nhiệm vụ phức tạp khác. Giờ đây ông đang khám phá cách kết nối điện toán các module đó với nhau.[15]

Máy học có lẽ là một khái niệm mới đối với Good, nhưng chắc ông đã gặp các giải thuật kiểu máy học khi thẩm định máy Perceptron cho IBM. Khi đó, khả năng đáng sợ của việc máy có thể học như người đã khiến Good nghĩ đến những hậu quả mà những người khác chưa từng tưởng tượng ra. Nếu một cỗ máy có thể tự làm nó thông minh hơn, thì cái máy đã thông minh hơn đó thậm chí sẽ còn giỏi hơn trong việc tự làm nó thông minh, và cứ thế.

Trong những năm 1960 dữ dội ấy, cho đến khi sáng tạo ra khái niệm *sự bùng nổ trí thông minh,* có lẽ ông đã nghĩ đến những loại vấn đề mà một cỗ máy thông minh có thể giải quyết. Không còn các tàu ngầm U-boat của Đức để đánh chìm, nhưng vẫn còn đó đối thủ Liên Xô, cuộc khủng hoảng tên lửa Cuba, vụ ám sát Tổng thống Kennedy, cuộc chiến ủy nhiệm giữa Mỹ và Trung Quốc ở Đông Nam Á. Con người đã ở trên bờ vực của sự

diệt vong – và dường như lại cần đến một Người khổng lồ mới. Trong *Speculations,* Good viết:

> [Nhà tiên phong trong lĩnh vực máy tính] B. V. Bowden phát biểu rằng… không có lý do gì để xây dựng một cỗ máy với trí thông minh của con người, vì sẽ dễ hơn nhiều nếu ta dùng cách thông thường…”◎ Câu này chứng tỏ một người rất thông minh cũng có thể bỏ qua cái gọi là “sự bùng nổ trí thông minh.” Về mặt kinh tế, đúng là sẽ lãng phí khi chế tạo một cỗ máy chỉ có khả năng làm những việc thông minh thông thường. Nhưng nếu điều này có thể thực hiện được, thì tương đối chắc chắn rằng nếu ta bỏ tiền ra gấp đôi, cái ta thu được sẽ là một cỗ máy cực thông minh.[16]

Vậy là thêm một ít đô-la nữa bạn sẽ có ASI, siêu trí tuệ nhân tạo, Good nói. Nhưng hãy coi chừng với những biến động lớn trong xã hội, khi ta chia sẻ hành tinh này với thứ thông minh hơn con người.

Vào năm 1962, trước khi viết *Speculations Concerning the First Ultraintelligent Machine*, Good là chủ biên một cuốn sách tên là *The Scientist Speculates* (Những suy đoán của các nhà khoa học). Ông viết một chương với nhan đề “The Social Implications of Artificial Intelligence” (Những ảnh hưởng xã hội của trí tuệ nhân tạo), một kiểu chuẩn bị cho ý tưởng về siêu trí tuệ nhân tạo mà ông đang phát triển. Giống như Steve Omohundro sẽ tranh luận vào 50 năm sau, ông lưu ý rằng trong những vấn đề mà máy móc thông minh sẽ phải giải quyết, có những thứ được tạo ra bởi chính sự xuất hiện gây chia rẽ của chúng trên Trái đất.

> Những máy móc như thế… thậm chí có thể đưa ra những lời khuyên hữu ích về chính trị và kinh tế; và chúng sẽ *cần phải* làm như vậy để bù đắp lại những vấn đề mà bản thân sự tồn tại của chúng gây ra. Sẽ có những vấn đề về quá tải dân số do bệnh tật sẽ bị đẩy lùi, và về thất nghiệp bởi các robot cấp thấp do máy móc thông minh chế tạo ra sẽ có hiệu suất cao hơn con người.[17]

Nhưng, như tôi đã sớm biết được, vào cuối đời Good đã thay đổi quan niệm một cách đáng ngạc nhiên. Tôi đã luôn gộp ông vào nhóm những người lạc quan như Ray Kurzweil, vì ông đã chứng kiến máy móc "cứu" thế giới trước đây, và bài tiểu luận của ông đã gắn sự sống còn của nhân loại với công cuộc chế tạo ra máy tính siêu thông minh. Nhưng bạn ông, Leslie Pendleton, đã nói bóng gió đến một sự thay đổi. Bà cần thời gian để nhớ lại lúc đó, và đến ngày cuối của tôi ở Blacksburg, bà đã nhớ ra.

Năm 1998, Good được trao Giải thưởng Nhà tiên phong Máy tính© của Cộng đồng Máy tính IEEE (Institue of Electrical and Electronics Engineers: Viện Kỹ sư điện và điện tử). Ông đã 82 tuổi. Ông được yêu cầu cung cấp một tiểu sử, như một phần của bài phát biểu nhận giải. Ông đã gửi nó, nhưng cũng như bất kỳ ai khác, ông không đọc nó trong lễ nhận giải. Có lẽ chỉ mình Pendleton biết nó tồn tại. Bà đã thêm một bản sao của nó vào cùng vài tài liệu khác mà tôi yêu cầu, và đưa cho tôi trước khi tôi rời Blacksburg.

Trước khi đi vào quốc lộ I-81 và hướng về phía Bắc, tôi đã đọc nó trên xe trong bãi đỗ thuộc trung tâm điện toán đám mây của công ty Rackspace. Cũng như Amazon và Google, Rackspace (khẩu hiệu công ty: *Fanatical Support,* nghĩa là hỗ trợ cuồng nhiệt) cung cấp sức mạnh điện toán cực lớn

với giá rẻ bằng cách cho thuê thời gian sử dụng hàng chục ngàn bộ vi xử lý và hàng triệu gigabyte dung lượng. Tất nhiên là Đại học Virginia Tech với khẩu hiệu "Sáng tạo tương lai" sẽ có quyền tiếp cận một cơ sở của Rackspace, và tôi thì muốn tham quan, tuy nhiên lúc đó nó đang đóng cửa. Chỉ sau đó tôi mới thấy có vẻ kỳ lạ khi chỉ cách vài chục mét kể từ chỗ tôi ngồi đọc bản tiểu sử của Good, hàng chục ngàn bộ vi xử lý được làm mát đang chạy hết công suất, giải quyết các vấn đề của thế giới.

Trong tiểu sử, với cách viết láu lỉnh ở ngôi thứ ba, Good tổng kết các cột mốc đời mình, bao gồm cả những điều chưa từng được kể về công việc của ông tại Bletchley Park cùng với Turing. Nhưng dưới đây là những gì ông viết vào năm 1998 về siêu trí tuệ nhân tạo đầu tiên, và về sự quay ngoắt thái độ của ông trong những năm tháng cuối đời:

> [Bài viết] "Speculations Concerning the First Ultraintelligent Machine" (1965)… bắt đầu: "Sự tồn tại của con người phụ thuộc vào những bước đầu tiên trong việc xây dựng một cỗ máy cực trí thông minh." Đó là những câu chữ của ông [Good] trong Chiến tranh Lạnh, và bây giờ ông ngờ rằng từ "tồn tại" nên được thay bằng "tuyệt chủng." Ông nghĩ rằng, do cuộc cạnh tranh quốc tế, chúng ta không thể ngăn cản máy móc đoạt lấy quyền lực. Ông nghĩ rằng chúng ta là những con chuột nhắt. Ông cũng từng nói: "Có lẽ Con người sẽ tạo nên một *vị chúa đến từ cỗ máy*◎, theo hình dáng của mình."[18]

Tôi đọc những dòng đó và nhìn trân trân đầy ngớ ngẩn vào tòa nhà Rackspace. Đến cuối đời, Good đã xét lại không chỉ niềm tin của ông về khả năng tồn tại của Chúa. Tôi đã tìm thấy thông điệp trong chai, một chú thích đã thay đổi tất cả. Giờ đây, Good và tôi đã có một điểm chung quan

trọng. Chúng tôi đều tin rằng sự bùng nổ trí thông minh sẽ không có một kết cục tốt.

8

Điểm Không Thể Quay Đầu

> Nhưng nếu Singularity – điểm kỳ dị của phát triển công nghệ có thể xảy ra, nó sẽ xảy ra. Thậm chí nếu tất cả các chính phủ trên thế giới nhận thức rõ "mối nguy" này và cực kỳ lo sợ về nó, thì tiến trình tới mục tiêu vẫn sẽ tiếp tục. Trên thực tế, lợi thế cạnh tranh – trong kinh tế, quân sự, kể cả nghệ thuật – của mọi tiến bộ trong tự động hóa đều hấp dẫn đến nỗi có đưa ra các đạo luật hoặc thuế khóa hòng ngăn cản chúng cũng chỉ đơn thuần là đưa cờ cho người khác phất.[1]
>
> — **Vernor Vinge**
> *The Coming Technological Singularity*
> (Điểm kỳ dị công nghệ sắp tới), 1993

Đoạn trích trên nghe giống như một phiên bản được cắt riêng ra từ tiểu sử của I. J. Good, phải vậy không? Giống như Good, Giáo sư toán học Vernor Vinge, tác giả truyện khoa học viễn tưởng hai lần đoạt Giải thưởng Hugo, đã mượn ý văn của Shakespeare để ám chỉ chuyện giống như lũ chuột nhắt, thói chuộng danh lợi của con người sẽ đút đầu họ vào nòng đại bác. Vinge nói với tôi, ông chưa bao giờ đọc các đoạn tiểu sử tự viết của

Good, hoặc biết đến chuyện ông ấy đã thay đổi quan niệm của mình về sự bùng nổ trí thông minh vào những năm cuối đời. Có lẽ chỉ có Good và Leslie Pendleton là biết chuyện đó.◎

Vernor Vinge là người đầu tiên chính thức dùng thuật ngữ Singularity để mô tả tương lai công nghệ – năm 1993, ông đã viết nó ra trong một bài gửi cho NASA có tên là "The Coming Technological Singularity" (Điểm kỳ dị công nghệ sắp tới). Nhà toán học Stanislaw Ulam thông báo rằng bà và học giả John von Neumann đã sử dụng "Singularity" trong một cuộc trò chuyện về sự thay đổi công nghệ từ 35 năm trước, năm 1958. Nhưng việc đặt ra từ mới này của Vinge là công khai, có chủ đích, và cổ vũ Ray Kurzweil bắt đầu dùng thuật ngữ này mà hiện nay đã trở thành phong trào.

Với uy tín như vậy, tại sao Vinge không giảng dạy và thuyết trình tại các hội thảo với tư cách chuyên gia hàng đầu về Singularity?

Thực ra Singularity có nhiều nghĩa, và cách dùng của Vinge chính xác hơn những người khác. Để định nghĩa Singularity, ông đã làm phép so sánh giữa nó với một điểm trên quỹ đạo của lỗ đen, nơi mà đi quá nó thì đến ánh sáng cũng không thoát ra nổi. Bạn không thể thấy những gì xảy ra bên kia điểm đó, được gọi là chân trời sự kiện. Tương tự, khi chúng ta chia sẻ hành tinh này với các thực thể thông minh hơn, mọi sự đánh cược đều vô nghĩa – chúng ta không thể đoán được điều gì sẽ xảy ra. Bạn cần phải thông minh ít nhất bằng chúng thì mới hiểu được.

Vậy, nếu bạn không có khái niệm rõ ràng gì về những thứ sẽ xảy ra trong tương lai, làm sao bạn có thể viết về nó? Vinge không viết truyện khoa học giả tưởng, ông có tiếng là một tác giả truyện khoa học viễn

tưởng, sử dụng các kiến thức khoa học có thật trong các tác phẩm của ông. Singularity đã khiến ông phải câm lặng.

> Đây là một vấn đề mà chúng ta luôn gặp khi nghĩ đến sự ra đời của trí thông minh vượt cấp độ con người. Khi điều đó xảy ra, lịch sử nhân loại sẽ chạm đến một thứ giống như điểm kỳ dị – một nơi mà các phép ngoại suy không còn ý nghĩa và những hình mẫu mới cần được áp dụng – và thế giới này sẽ vượt quá tầm hiểu biết của chúng ta.[3]

Như Vinge đã nói, khi ông bắt đầu viết truyện khoa học viễn tưởng vào những năm 1960, những thế giới mà ông tưởng tượng ra trên cơ sở khoa học nằm trong khoảng 40 đến 50 năm sau. Nhưng đến những năm 1990, tương lai chạy *về phía* ông, và tiến bộ công nghệ dường như đang tăng tốc. Ông không còn có thể tiên liệu tương lai sẽ mang đến những gì, vì ông cho rằng trí thông minh vượt cấp độ con người sẽ sớm xuất hiện. Trí thông minh đó, chứ không phải của con người, sẽ là thứ quyết định tốc độ phát triển công nghệ. Ông không thể viết về nó, và những người khác cũng vậy.

> Qua những năm 60, 70 rồi 80, ý tưởng về tận thế được chấp nhận rộng rãi. Có lẽ các tác gia khoa học viễn tưởng là những người đầu tiên cảm thấy điều này một cách rõ rệt nhất. Nói cho cùng thì các tác gia khoa học viễn tưởng "chân chính" này là những người cố gắng viết những câu chuyện cụ thể về những gì mà công nghệ sẽ mang đến cho chúng ta. Các tác gia này ngày càng cảm thấy có một bức tường mờ đục chắn ngang tương lai.[4]

• • •

Nhà nghiên cứu AI Ben Goertzel đã nói với tôi: "Vernor Vinge nhìn thấy tính bất khả tri cố hữu của nó rất rõ ràng khi ông đưa ra khái niệm điểm kỳ dị công nghệ. Vì thế mà ông không thuyết giảng vòng vo về nó, bởi ông không biết phải nói gì. Ông có thể nói gì được? 'Vâng, tôi nghĩ rằng chúng ta sẽ tạo ra các công nghệ có khả năng vượt trội hơn con người rất nhiều, và sau đó ai mà biết được chuyện gì sẽ xảy ra?' "

Nhưng những phát minh ra lửa, nông nghiệp, in ấn và điện thì sao? Nhiều "Singularity" công nghệ phải chăng đã diễn ra rồi? Sự thay đổi lớn về công nghệ không phải là điều gì mới, không ai cảm thấy cần phải nghĩ ra một cái tên mỹ miều cho chúng. Bà của tôi sinh ra trước khi xe ô tô được sử dụng rộng rãi, và bà đã chứng kiến Neil Armstrong bước đi trên Mặt trăng. Bà đặt tên cho sự thay đổi ấy là thế kỷ 20. Vậy có gì đặc biệt đến thế trong sự chuyển tiếp mà Vinge đề cập?

"Yếu tố bí mật ở đây là trí thông minh," Vinge nói với tôi. Ông có giọng nói liến thoắng với âm vực tenor nam cao, khiến người ta buồn cười. "Trí tuệ là thứ tạo nên sự khác biệt, và đặc trưng để phân biệt là những sinh vật ở trước thời điểm đó sẽ chẳng thể hiểu được. Chúng ta đang ở tình thế là trong khoảng thời gian rất ngắn, chỉ vài thập niên, chúng ta sẽ thấy những sự biến đổi tương đương với những tiến hóa sinh học lớn."

Trong câu trên có hai ý quan trọng. Thứ nhất, Singularity công nghệ sẽ mang tới một sự thay đổi trong bản thân trí thông minh, siêu sức mạnh duy nhất của con người, thứ đã khởi tạo nên công nghệ. Đó là lý do vì sao nó khác với bất cứ cuộc cách mạng nào. Thứ hai, sự tiến hóa sinh học mà Vinge nói đến là khi loài người chiếm lĩnh hành tinh này vào 200.000 năm trước. Vì họ thông minh hơn muôn loài, nên *Homo sapiens*, hay "người

tinh khôn," bắt đầu thống trị hành tinh. Tương tự, những ý thức một ngàn hoặc một triệu lần thông minh hơn con người sẽ thay đổi cuộc chơi này mãi mãi. Và điều gì sẽ xảy ra với chúng ta?

Câu hỏi này khiến Vinge cười khùng khục: "Nếu tôi bị hỏi dồn về những thứ xảy ra sau Singularity, cách thoái thác tôi hay dùng nhất là nói bạn nghĩ tại sao tôi lại đặt tên nó là điểm kỳ dị?"

Nhưng Vinge đã kết luận được một điều về cái tương lai bất định ấy – Singularity là đáng sợ, và có thể dẫn tới sự diệt vong của chúng ta. Trong bài phát biểu vào năm 1993, ông đã trích dẫn đoạn nói về sự bùng nổ trí thông minh trong bài báo của Good năm 1967, chỉ ra rằng nhà thống kê học nổi tiếng này phân tích còn chưa đủ sâu:

> Good đã nắm bắt được về bản chất của việc AI chạy trốn, nhưng không khảo sát những hệ quả khủng khiếp nhất của nó. Bất cứ cỗ máy thông minh nào kiểu như ông đã mô tả sẽ không phải là "công cụ" của con người – chẳng khác gì con người không phải là công cụ của loài thỏ, chim cổ đỏ hoặc tinh tinh.[5]

Đây là một sự so sánh đúng đắn – thỏ đối với người cũng như người đối với máy móc siêu thông minh. Chúng ta đối xử với loài thỏ thế nào? Như một thứ động vật phá hoại, thú cưng, hoặc bữa tối. Các ASI lúc đầu sẽ là công cụ của chúng ta – giống các tiền bối của chúng hiện tại như Google, Siri và Watson. Và Vinge đề xuất thêm rằng có nhiều cách khiến Singularity xảy ra bên cạnh trí thông minh máy tính đơn thuần. Chúng bao gồm trí thông minh lớn lên từ Internet, từ Internet *cộng thêm* những người dùng (một Gaia◎ kỹ thuật số), từ giao diện người-máy, và từ khoa học sinh

học (tăng cường trí thông minh của các thế hệ tương lai qua việc điều khiển gen).[6]

Ở ba trong số những cách trên, con người tham dự xuyên suốt quá trình phát triển công nghệ, có lẽ lèo lái một quá trình vươn lên trong trí thông minh một cách tuần tự và kiểm soát được, chứ không phải là một *vụ bùng nổ*. Vinge nói, nghĩa là có thể những trở ngại to lớn nhất của loài người – nạn đói, bệnh tật, kể cả cái chết – sẽ bị chinh phục. Đó là một viễn cảnh được Ray Kurzweil tán thành, được "những người theo thuyết Singularity" truyền bá. Những người theo thuyết Singularity là những người tiên đoán rằng sẽ chỉ có hầu như toàn điều tốt đẹp trong cái tương lai được tăng tốc sắp tới. Singularity của họ nghe có vẻ quá lạc quan đối với Vinge.

"Chúng ta đang chơi một trò chơi cực kỳ nguy hiểm, và cái mặt tốt của nó thì lại lạc quan đến mức đáng sợ. Một làn gió mới trong nền kinh tế thế giới được gắn liền với sự phát triển AI. Và đó là một nguồn lực vô cùng mạnh mẽ. Do đó, hàng trăm ngàn người trên thế giới, rất thông minh, đang làm những thứ sẽ dẫn tới trí thông minh siêu nhân. Và nhiều khả năng hầu hết họ thậm chí không nhìn nó theo cách đó. Họ chỉ thấy những thứ như là *nhanh hơn, rẻ hơn, tốt hơn, nhiều lợi nhuận hơn?*

Vinge so sánh nó với một chiến lược trong Chiến tranh Lạnh gọi là MAD (Mutually Assured Destruction: đảm bảo tiêu diệt lẫn nhau). Được người thích các loại tên rút gọn là John von Neumann (cũng là người sáng chế một trong những máy tính đầu tiên với tên rút gọn MANIAC) đặt tên, MAD đã duy trì hòa bình trong Chiến tranh Lạnh bằng viễn cảnh hai bên cùng bị xóa sổ. Cũng như MAD, trong lĩnh vực siêu trí tuệ nhân tạo hiện có nhiều nhà nghiên cứu đang bí mật làm việc để phát triển những công nghệ có tiềm năng gây ra thảm họa. Nhưng khác với chiến lược MAD, ở đây

không có điểm dừng nào cả. Không ai biết ai đang dẫn đầu, vì thế mọi người đều giả định rằng có ai đó đang đi trước. Và như chúng ta đã thấy, người thắng sẽ không có tất cả. Người thắng trong cuộc chạy đua vũ trang AI này sẽ chỉ giành được cái hư danh của người đầu tiên đương đầu với Đứa trẻ Bận rộn, hoặc cố gắng sống sót trong cái kịch bản mà nó tạo ra.

"Chúng ta đang có hàng ngàn người giỏi làm việc khắp nơi trên thế giới theo kiểu cùng chung sức để gây ra một thảm kịch," Vinge nói. "Viễn cảnh về các mối đe dọa sắp tới là rất xấu. Chúng ta không chịu gắng sức suy nghĩ về những khả năng thất bại."

• • •

Một số kịch bản khác mà Vinge quan tâm cũng đòi hỏi nhiều sự chú ý hơn. Một Gaia kỹ thuật số, hay sự kết hợp giữa mạng lưới người dùng và máy tính, hiện đã bắt đầu hình thành trên Internet. Ý nghĩa của điều đó đối với tương lai chúng ta là sâu sắc và rộng lớn, đáng để có nhiều cuốn sách hơn viết về nó. IA (Intelligence Augmentation: sự tăng cường trí thông minh), có tiềm năng dẫn đến tai họa tương tự như AI thuần, với đôi chút giảm nhẹ khi có sự góp mặt của con người, chí ít là ở các giai đoạn đầu. Nhưng ưu thế đó sẽ nhanh chóng biến mất. Chúng ta sẽ nói về IA sau. Bây giờ tôi muốn chú ý tới ý tưởng của Vinge rằng trí thông minh có thể khởi sinh từ Internet.

Một số nhà tư tưởng công nghệ, trong đó có George Dyson và Kevin Kelly, đã đề xuất rằng thông tin là một dạng sống. Mã máy tính mang thông tin tự nhân bản và lớn lên theo những quy luật sinh học.[7] Nhưng với trí thông minh, đó lại là một câu chuyện khác. Nó là một đặc điểm của các tổ chức sinh học phức tạp, và nó không tự nhiên sinh ra.

Tại ngôi nhà của ông ở California, tôi đã hỏi Eliezer Yudkowsky rằng liệu trí thông minh có thể sinh ra từ phần cứng của Internet đang phát triển không ngừng với hàm mũ, từ cơ sở dữ liệu tương đương năm triệu triệu megabyte, từ mạng được kết nối bởi hơn bảy tỉ máy tính và điện thoại thông minh, và từ 75 triệu máy chủ của nó.[8] Yudkowsky đã cau mày, như thể các tế bào não của ông vừa bị nhấn chìm trong sự câm lặng.

"Hoàn toàn không," ông nói. "Phải mất hàng tỉ năm để quá trình tiến hóa có thể đẻ ra trí thông minh. Nó không sinh ra từ sự phức tạp của sự sống. Nó không tự động xảy đến. Cần có áp lực tối ưu hóa của chọn lọc tự nhiên."

Nói cách khác, trí thông minh không nảy sinh từ sự phức tạp đơn thuần. Và Internet thì thiếu kiểu áp lực môi trường vốn ưu đãi đột biến này hơn đột biến khác trong tự nhiên.

"Tôi có một cách nói rằng có lẽ sự phức tạp thú vị trong cả giải Ngân hà ngoài kia còn ít hơn cả trong một con bướm nhỏ, vì con bướm được tạo ra từ những tiến trình đã giữ lại các thành tựu và tiếp tục xây dựng trên đó," Yudkowsky nói.

Tôi đồng ý rằng trí thông minh sẽ không tự nhiên nở rộ trên Internet. Nhưng tôi nghĩ mô hình tài chính tác tử (agent-based financial modeling) có thể sẽ sớm thay đổi mọi thứ về Internet.

Ngày trước, khi các nhà phân tích phố Wall muốn dự đoán xem thị trường hành xử như thế nào, họ quay về với một loạt những quy luật được kinh tế học vĩ mô vạch ra. Những quy luật đó xem xét tới các yếu tố như lãi suất, dữ liệu việc làm, các vấn đề nhà ở như bao nhiêu nhà mới được xây. Tuy nhiên, phố Wall ngày càng chuyển sang dùng mô hình tài chính

tác tử. Ngành khoa học mới này có thể mô phỏng điện toán toàn bộ thị trường chứng khoán và kể cả nền kinh tế, hòng tăng cường khả năng dự báo.

Để mô phỏng thị trường, các nhà nghiên cứu tạo ra các mẫu trong máy tính về những thực thể mua và bán chứng khoán – các cá nhân, công ty, ngân hàng, quỹ đầu tư… Mỗi loại có hàng ngàn "tác tử" với những mục đích, quy tắc hành xử, chiến thuật để mua và bán khác nhau. Chúng lại bị dữ liệu thị trường liên tục thay đổi tác động. Các tác tử này, do mạng neuron nhân tạo và các kỹ thuật AI khác vận hành, được "huấn luyện" trên các thông tin đời thực. Hoạt động đồng bộ và được cập nhật dữ liệu tức thời, các tác tử tạo ra một chân dung luôn dịch chuyển của thị trường chứng khoán sôi động.

Sau đó, các nhà phân tích thử nghiệm các kịch bản cho việc giao dịch một số chứng khoán nào đó. Và thông qua các kỹ thuật lập trình tiến hóa, mô hình thị trường này có thể "ngoại suy" ra một ngày hoặc một tuần, cung cấp cho các nha phân tích môt cái nhìn đúng về thị trường trong tương lai, và những cơ hội đầu tư nào sẽ xuất hiện. Cách tiếp cận "từ dưới lên" với việc xây dựng các mô hình tài chính này biểu thị ý tưởng là những quy tắc hành xử đơn giản của các tác nhân đơn lẻ sẽ gộp lại thành hành xử phức tạp chung. Nói một cách tổng quát, cái đúng với tổ ong và tổ kiến cũng sẽ đúng với phố Wall.

Thứ bắt đầu định hình trong các siêu máy tính ở những trung tâm tài chính thế giới là các thế giới ảo ngập tràn các chi tiết của thế giới thực, và cư trú ở đó là các "tác tử" ngày càng thông minh. Dự báo được nhiều điều hơn, sâu sắc hơn đồng nghĩa với lợi nhuận lớn hơn. Vậy là những lợi ích

kinh tế to lớn đã thúc đẩy mong muốn tăng cường độ chính xác của các mô hình ở mọi mức độ.

Nếu việc tạo ra các tác tử điện toán, vốn thực hiện các chiến thuật mua bán chứng khoán phức tạp, là hữu dụng, thì phải chăng việc tạo ra các mô hình điện toán có đầy đủ những động cơ và kỹ năng của con người còn có ích *hơn?* Tại sao không xây dựng AGI, hay các tác tử với trí thông minh cấp độ con người? Ồ, đấy chính là thứ mà phố Wall đang làm, nhưng với một cái tên khác – các mô hình tài chính tác tử.

Thị trường chứng khoán sẽ sinh ra AGI, đó là luận điểm của Tiến sĩ Alexander D. Wissner-Gross, người có lý lịch khiến các nhà phát minh, học giả và nhà thông thái khác phải nghiêng mình thán phục. Ông là tác giả của 13 cuốn sách, nắm giữ 16 bằng sáng chế, đã học đồng thời ba chuyên ngành Vật lý, Khoa học điện tử và kỹ thuật, Toán học tại MIT, tốt nghiệp thủ khoa School of Engineering của MIT. Ông có bằng tiến sĩ vật lý từ Harvard với luận án giành giải thưởng lớn. Ông đã thành lập và bán nhiều công ty, và theo bản trích ngang của ông, đã giành "107 danh hiệu lớn" mà có lẽ không cái nào trong đó thuộc kiểu xô bồ "nhân viên của tuần." Ông hiện là một nhà nghiên cứu của Harvard, đang cố gắng thương mại hóa ý tưởng của ông về tài chính điện toán.

Ông nghĩ rằng trong khi các nhà lý thuyết lỗi lạc khắp thế giới đang chạy đua để chế tạo AGI, nó có thể xuất hiện ở dạng hoàn chỉnh trong các thị trường tài chính, như là một hệ quả không lường trước của việc xây dựng các mô hình điện toán về những đám đông. Ai sẽ tạo ra nó? "Quant," thuật ngữ chỉ những chuyên gia tin học tài chính của phố Wall.

“Hoàn toàn có khả năng một AGI đầy sinh động sẽ trỗi dậy từ các thị trường tài chính,” Wissner-Gross nói với tôi. “Không phải là từ kết quả một giải thuật đơn nhất của một quant đơn lẻ, mà từ một tổng thể của tất cả các thuật toán đến từ rất nhiều quỹ đầu tư. AGI có thể không cần một lý thuyết chặt chẽ. Nó có thể là một hiện tượng kết tụ. Nếu người ta còn thích tiền, thì ngành tài chính còn nhiều khả năng là bồn nước nguyên thủy mà từ đó sự sống của AGI sẽ nảy mầm.”

Để kịch bản này xảy ra, cần phải có thật nhiều tiền rót vào việc chế tạo các mô hình tài chính ngày càng siêu việt hơn. Và thực tế đúng như vậy – tin đồn là tiền đổ vào trí thông minh máy móc hay AGI ở đây còn nhiều hơn bất cứ chỗ nào, có lẽ còn hơn cả DARPA, IBM và Google. Điều đó sẽ chuyển hóa thành nhiều siêu máy tính mạnh hơn và các quant giỏi hơn. Wissner-Gross nói rằng các quant sử dụng cùng một bộ công cụ với các nhà nghiên cứu AI – mạng neuron, giải thuật di truyền, đọc tự động, các mô hình Markov ẩn… Mọi công cụ AI mới đều được kiểm tra trong lò lửa tài chính.

“Bất cứ khi nào một kỹ thuật AI mới được phát triển,” Wissner-Gross nói với tôi, “câu đầu tiên từ miệng ai đó sẽ là có dùng để chơi chứng khoán được không?’”

Bây giờ hãy tưởng tượng bạn là một quant có thế lực, với một ngân sách đủ lớn để thuê các quant khác và mua thêm phần cứng. Quỹ đầu tư mà bạn làm việc đang chạy một mô hình lớn mô phỏng phố Wall, gồm hàng ngàn các tác tử kinh tế đủ kiểu. Các giải thuật của nó tương tác với các giải thuật của các quỹ đầu tư khác – chúng gắn kết chặt chẽ đến nỗi chúng lên xuống cùng nhau như là đang cùng diễn tấu trong buổi hòa nhạc. Theo Wissner-Gross, những nhà quan sát thị trường đã đưa ra giả thuyết là một

số dường như đã *ra hiệu* cho nhau với những giao dịch cỡ mili giây, xảy ra với tốc độ mà con người không thể theo kịp (đây là HFT hay các giao dịch tần số cao, đã được thảo luận ở Chương 6).

Phải chăng bước đi logic tiếp theo là làm cho quỹ đầu tư của bạn biết suy nghĩ? Nghĩa là, có lẽ giải thuật của bạn không nên tự động khởi phát các lệnh bán dựa trên các lệnh bán tháo ồạt của một quỹ đầu tư khác (điều đã diễn ra vào lần Sụp đổ Chớp giật vào tháng 5/2010). Thay vào đó, nó sẽ nhận thức việc bán tháo và quan sát xem điều đó đang ảnh hưởng đến các quỹ khác và toàn thị trường thế nào, trước khi làm điều gì đó. Nó có thể phản ứng một cách khác biệt, tốt hơn. Hoặc nó có thể làm hơn thế, giả lập đồng thời một số cực lớn các thị trường giả định, và sẵn sàng thực hiện một trong nhiều chiến thuật tương ứng với các tình thế nhất định.

Nói cách khác, sẽ có những lợi ích tài chính khổng lồ nếu giải thuật của bạn có khả năng tự nhận thức – để biết chính xác nó là gì và mô phỏng thế giới xung quanh nó. Cái này *rất giống* AGI. Đây chính xác là cách thị trường đang theo đuổi, nhưng liệu có ai đó đang đi tắt đón đầu và chế tạo AGI?

Wissner-Gross không biết. Và có lẽ ông sẽ không nói với bạn nếu ông biết. "Có sự thôi thúc mạnh mẽ kiểu nhà buôn khiến bạn muốn giữ bí mật về bất cứ thứ gì mang lại lợi nhuận lớn," ông nói.

Tất nhiên. Và ông không chỉ nói đến sự cạnh tranh giữa các quỹ đầu tư, mà cả kiểu chọn lọc tự nhiên trong số các giải thuật. Người thắng cuộc sẽ sống và giữ được mã của mình. Kẻ thất bại sẽ chết. Áp lực tiến hóa của thị trường sẽ làm tăng tốc sự phát triển của trí thông minh, nhưng không thể không có sự chỉ dẫn của các quant con người. Chí ít là đến lúc này.

Một sự bùng nổ trí thông minh sẽ khó nhận ra trong thế giới tài chính điện toán, vì ít nhất là bốn lý do. Thứ nhất, như nhiều kiến trúc nhận thức, nó có khả năng sẽ sử dụng mạng neuron, lập trình di truyền, và những kỹ thuật AI "hộp đen" khác. Thứ hai, các cuộc truyền tín hiệu với băng thông rộng và tốc độ cỡ mili giây đã vượt cấp độ phản ứng của con người – hãy xem điều gì đã xảy ra trong Sụp đổ Chớp giật. Thứ ba, hệ thống này vô cùng phức tạp – không có quant hoặc thậm chí một nhóm quant nào có khả năng giải thích hệ sinh thái của các giải thuật ở phố Wall, và chuyện các giải thuật tương tác ra sao.

Cuối cùng, nếu một trí thông minh ghê gớm nảy sinh từ tài chính điện toán, nó sẽ hầu như chắc chắn được giữ bí mật chừng nào nó còn sinh lời cho những người tạo ra nó. Đấy là bốn mức độ của sự thiếu minh bạch.

Tổng kết lại, AGI có thể trỗi dậy từ phố Wall. Những giải thuật thành công nhất được giữ bí mật bởi các quant tạo ra chúng một cách đầy ưu ái, hoặc những công ty sở hữu chúng. Một sự bùng nổ trí thông minh sẽ là không thể nhận thấy đối với hầu hết, nếu không muốn nói là cả nhân loại, và dù sao thì đó dường như cũng là chuyện không thể ngăn chặn.

Những sự tương đồng giữa tài chính điện toán và nghiên cứu AGI không dừng lại ở đó. Wissner-Gross có một đề xuất lạ lùng khác. Ông cho rằng những chiến thuật đầu tiên để kiểm soát AGI có thể nảy sinh từ những cách thức hiện đang được đề xuất để kiểm soát giao dịch tần số cao. Một số trong đó có vẻ khá triển vọng.

Câu dao thị trường sẽ cắt những AI của các quỹ đầu tư ra khỏi thế giới bên ngoài, trong trường hợp khẩn cấp. Chúng sẽ dò ra các tương tác ghép

tầng của các giải thuật như trường hợp Sụp đổ Chớp giật năm 2010 và sẽ ngắt kết nối máy tính.

Luật Giao dịch Lớn đòi hỏi AI phải được đăng ký chi tiết, cùng với biểu đồ tổ chức nhân sự đầy đủ. Nếu điều này nghe như là tiền đề của sự can thiệp lớn từ chính quyền, thì đúng thế. Tại sao không? Phố Wall đã chứng minh hết lần này đến lần khác rằng trong vai trò của một nền văn hóa, nó không thể cư xử có trách nhiệm nếu không có quy định nghiêm khắc. Điều này có đúng với cả các nhà phát triển AGI không? Chẳng nghi ngờ gì nữa. Nào có cần các huy hiệu đạo đức để nghiên cứu AGI.

Kiểm tra giải thuật tiên giao dịch có thể giả lập các hành vi của giải thuật trong một môi trường ảo, trước khi chúng được thả ra thị trường. *Bài kiểm tra Mã nguồn AI và Biên bản tập trung các hoạt động AI* nhắm đến việc lường trước các sai sót, trợ giúp việc phân tích sau khi các tai nạn xảy ra, kiểu như vụ Sụp đổ Chớp giật năm 2010.

Nhưng hãy nhìn lại bốn mức độ của sự thiếu minh bạch trên đây, và hãy xem liệu những biện pháp phòng thủ này, ngay cả khi chúng được triển khai đầy đủ, nghe có giống một sự bảo đảm chắc chắn với bạn không.

• • •

Như chúng ta đã thấy, Vinge đã tiếp nối I. J. Good và đưa vào sự bùng nổ trí thông minh những thuộc tính mới quan trọng. Ông xem xét những con đường khác để đạt tới nó ngoài phương pháp mạng neuron của Good, và chỉ ra những khả năng, thậm chí xác suất, của việc con người bị xóa sổ. Và có lẽ điều quan trọng nhất là Vinge đã cho nó một cái tên – Singularity.

Đặt tên cho sự vật, điều mà Vinge, tác giả của tiểu thuyết khoa học viễn tưởng *True Names* (Những cái tên thật) biết rất rõ, là một hành động đầy sức mạnh. Những cái tên gắn nơi đầu môi, lưu lại trong bộ não, và truyền miệng qua các thế hệ. Những nhà thần học đề xuất trong sách Sáng thế ký, rằng việc đặt tên mọi sự vật trên Trái đất vào ngày thứ bảy là quan trọng, bởi một tạo vật duy lý sẽ chia sẻ sân khấu do Chúa tạo ra, và sau đó sẽ cần dùng những cái tên. Phát triển từ vựng là một giai đoạn quan trọng trong sự phát triển của trẻ em. Chúng ta biết rằng không có ngôn ngữ, bộ não không thể phát triển bình thường. Dường như AGI sẽ khó thành hiện thực nếu không có ngôn ngữ, không có các danh từ, không có những cái tên.

Vinge đặt tên Singularity để chỉ một thời điểm đáng sợ đối với con người, một kế hoạch không an toàn. Định nghĩa của ông về Singularity có tính ẩn dụ – quỹ đạo ngoài lỗ đen, nơi lực hấp dẫn mạnh đến nỗi ngay cả ánh sáng cũng không thoát ra được. Chúng ta không thể biết bản chất của nó, và nó được đặt tên như vậy là có chủ đích.

Rồi đột nhiên, mọi thứ thay đổi.

Dựa trên ý tưởng về Singularity của Vinge, Ray Kurzweil đã thêm vào một chất xúc tác đầy kịch tính, đưa cả cuộc đàm luận này lên bệ phóng, và tai họa sắp tới trở nên rõ ràng hơn: sự tăng trưởng theo hàm mũ của sức mạnh và tốc độ máy tính. Chính vì sự tăng trưởng này mà những người cho rằng máy tính có trí thông minh cấp độ con người là bất khả thi trong thế kỷ này hoặc về sau nói chung, sẽ bị những người khác nhìn với ánh mắt nghi hoặc.

Với mỗi đô-la được tiêu◎, sức mạnh của máy tính đã tăng lên cả tỉ lần trong 30 năm vừa qua. Sau 20 năm nữa, 1.000 đô-la sẽ mua được chiếc

máy tính mạnh hơn chiếc hôm nay cả triệu lần, và sau 25 năm thì cả *tỉ* lần rnạnh hơn chiếc hôm nay. Vào khoảng năm 2020, máy tính sẽ có thể mô phỏng não người, và vào khoảng năm 2029, các nhà nghiên cứu sẽ có thể chạy một chương trình giả lập bộ não với trí thông minh và cảm xúc tinh tế như não người. Vào năm 2045, trí thông minh của người và máy sẽ tăng lên cả *tỉ lần,* sẽ phát triển những công nghệ đánh bại các điểm yếu của con người như mệt mỏi, bệnh tật và cái chết. Trong trường hợp chúng ta còn sống sót, thế kỷ 21 sẽ không chứng kiến một sự phát triển công nghệ tương đương với 100 năm, mà là 200.000 năm.[9]

Những phân tích và dự báo táo bạo này là của Kurzweil, nó là chìa khóa để hiểu định nghĩa thứ ba và bao trùm về khái niệm Singularity của Kurzweil. Nó ở trung tâm Định luật Hồi quy Tăng tốc của Kurzweil, một lý thuyết về sự tiến bộ công nghệ mà Kurzweil không phát minh ra, nhưng đã chỉ ra, tương đối giống với việc Good tiên đoán về sự bùng nổ trí thông minh và Vinge cảnh báo về Singularity đang tới. Định luật Hồi quy Tăng tốc đồng nghĩa với việc những dự báo và tiến bộ mà chúng ta đang thảo luận trong cuốn sách này sẽ chạy xình xịch hướng tới chúng ta như chuyến tàu hàng với tốc độ liên tục tăng lên gấp đôi sau mỗi dặm đường. Rất khó để nhận ra nó sẽ đến đây nhanh thế nào, nhưng đủ để nói rằng nếu con tàu đó đang chạy với tốc độ 20 dặm một giờ vào cuối dặm thứ nhất, thì chỉ 15 dặm sau nó đã chạy với tốc độ 65.000 dặm một giờ. Và điều quan trọng cần lưu ý là dự báo của Kurzweil không chỉ là về sự tiến bộ trong phần cứng công nghệ, giống như những thứ bên trong chiếc iPhone mới, mà về cả các tiến bộ trong nghệ thuật công nghệ, giống như việc phát triển một lý thuyết thống nhất về trí tuệ nhân tạo.

Nhưng Kurzweil và tôi chỉ có chung quan điểm đến đây. Thay vì dẫn tới thiên đường như các dự báo của Kurzweil khẳng định, tôi tin rằng Định luật Hồi quy Tăng tốc mô tả khoảng cách ngắn nhất có thể giữa cuộc đời chúng ta với sự kết thúc của kỷ nguyên con người.

9

Định Luật Hồi Quy Tăng Tốc

> Điện toán đang trải qua một sự thay đổi phi thường nhất kể từ thời điểm phát minh ra máy tính cá nhân. Những sáng tạo của thập niên sắp tới sẽ vượt xa những sáng tạo của cả ba thập niên trước gộp lại.[1]
>
> — **Paul Otellini**
> CEO của Intel

Với hai tác phẩm *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* và *The Singularity is Near,* Ray Kurzweil đã trưng dụng thuật ngữ Singularity và thay đổi ý nghĩa của nó thành một thời kỳ xán lạn đáng mong chờ trong lịch sử loài người, được ông quan sát với một độ chính xác đặc biệt qua các phép ngoại suy. Ông viết, sau khoảng 40 năm nữa, sự phát triển công nghệ sẽ tiến bộ nhanh đến mức sự hiện hữu của con người sẽ bị thay đổi về cơ bản, và kết cấu lịch sử vốn có sẽ sụp đổ. Máy móc và sinh vật sẽ trở nên không thể phân biệt được. Các thế giới ảo sẽ trở nên sinh động và quyến rũ hơn cả thế giới thực. Công nghệ nano sẽ cho phép sản xuất theo nhu cầu, chấm dứt nạn đói và sự nghèo khổ, chữa trị cho mọi loại bệnh tật của loài người. Bạn có thể dừng quá trình lão

hóa của mình lại, hoặc thậm chí đảo ngược nó. Đây sẽ là khoảng thời gian quan trọng nhất để sống, không chỉ vì bạn sẽ chứng kiến một nhịp độ thực sự đáng kinh ngạc của sự biến đổi công nghệ, mà còn bởi công nghệ hứa hẹn đem lại cho bạn các công cụ để trường sinh bất lão. Đó sẽ là bình minh của kỷ nguyên "Singularity."

> Vậy Singularity là gì? Đó là một giai đoạn tương lai, khi mà nhịp độ thay đổi công nghệ sẽ nhanh chóng và ảnh hưởng của nó sẽ sâu sắc đến nỗi cuộc sống con người sẽ bị biến đổi đến mức không thể đảo ngược. Dù không phải thiên đàng hay địa ngục, nhưng kỷ nguyên này sẽ biến đổi những khái niệm về đời sống mà chúng ta vẫn dựa vào, từ những mô hình kinh tế cho đến vòng quay của đời người, bao gồm cả bản thân cái chết…[2]
>
> Hãy xem những câu chuyện về Harry Potter của J. K. Rowling từ góc nhìn này. Những câu chuyện này có thể là tưởng tượng, nhưng nó không phải viễn cảnh vô lý vì thế giới của chúng ta sẽ có thể là như thế chỉ sau vài thập niên nữa tính từ bây giờ. Về cơ bản, mọi "phép thuật" của Potter sẽ được hiện thực hóa thông qua các công nghệ mà tôi sẽ khảo sát trong cuốn sách này. Trò quidditch và việc biến con người hay đồ vật thành những hình dạng khác sẽ có thể thực hiện được ◎ trong các môi trường giả lập hiện thực hoàn toàn, cũng như trong chính hiện thực bằng cách sử dụng các thiết bị nano.[3]

Vậy là thời kỳ Singularity sẽ "không phải là thiên đàng hay địa ngục" nhưng chúng ta sẽ được chơi quidditch! Khái niệm Singularity của Kurzweil hiển nhiên là rất khác biệt với khái niệm Singularity của Vernor Vinge và sự bùng nổ trí thông minh của I. J. Good. Liệu chúng có thể tương

thích? Đây vừa là quãng thời gian tốt nhất để sống, vừa là quãng thời gian tồi tệ nhất? Tôi đã đọc gần như từng từ mà Kurzweil viết, nghe tất cả các bài phát biểu, các đoạn ghi âm ghi hình của ông. Năm 1999, tôi đã phỏng vấn ông khá lâu cho bộ phim tài liệu có một phần về AI. Tôi biết ông đã viết và nói những gì về các mối nguy của AI, và những thứ đó không nhiều lắm.

Tuy vậy, đáng ngạc nhiên là ông lại chịu trách nhiệm gián tiếp cho sự ra đời của bài tiểu luận có tính cảnh báo thuyết phục nhất về Singularity – *Why the Future Doesn't Need Us* (1999) (Tại sao tương lai không cần chúng ta) của Bill Joy.[4] Trong đó Joy, một nhà lập trình, kiến trúc sư máy tính, đồng sáng lập Sun Microsystems, đã chủ trương cần giảm tốc và thậm chí dừng việc phát triển ba công nghệ mà ông tin là nguy hiểm chết người nếu cứ theo đuổi chúng với tốc độ hiện thời: trí tuệ nhân tạo, công nghệ nano và công nghệ sinh học. Joy cảm thấy bị thôi thúc phải viết tiểu luận này sau cuộc nói chuyện đáng sợ trong một quán bar với Kurzweil, sau khi đọc cuốn *The Age of Spiritual Machines* của ông. Trong các tác phẩm và bài giảng phi học thuật về hiểm họa AI, tôi nghĩ chỉ có Ba định luật của Asimov (dù nhầm lẫn) là được trích dẫn nhiều hơn bài luận có ảnh hưởng cực lớn của Joy. Đoạn văn sau sẽ tóm tắt cách nhìn của ông về AI:

> Nhưng, với viễn cảnh về sức mạnh máy tính cấp độ con người sau khoảng 30 năm nữa, một ý tưởng mới tự ra đời: phải chăng tôi đang làm việc để tạo ra những công cụ, thứ sẽ cho phép xây dựng công nghệ có khả năng thay thế loài người. Tôi cảm thấy sao về chuyện này? Rất không thoải mái. Trong cả sự nghiệp, tôi đã luôn đấu tranh để xây dựng các hệ thống phần mềm đáng tin cậy, và tôi thấy dường như chắc chắn rằng tương lai sẽ không tiếp diễn tốt đẹp như một số người

> đang mường tượng. Kinh nghiệm cá nhân cho tôi thấy chúng ta thường tự đánh giá quá cao khả năng thiết kế của mình. Trước sức mạnh khủng khiếp của những công nghệ mới này, chúng ta có nên đặt câu hỏi bằng cách nào chúng ta có thể sống chung với chúng? Và nếu từ những phát triển công nghệ ấy, chúng ta có thể hoặc thậm chí chắc chắn tuyệt chủng, thì chúng ta có nên tiếp tục với một sự cẩn trọng lớn?[5]

Vài lời nói chơi ở quán bar của Kurzweil có thể tạo ra một cuộc tranh luận tầm cỡ quốc gia, thế nhưng những lời cảnh báo ít ỏi của ông đã bị át đi trong cơn bão hứng khởi của những dự đoán tốt đẹp. Ông nhấn mạnh rằng mình không vẽ ra một ngày mai thiên đường, nhưng tôi nghĩ chắc chắn đó là thứ ông muốn làm.

Ít người viết một cách thông tuệ và thuyết phục về công nghệ như Kurzweil – ông kỳ công để làm cho thông điệp của mình thực sự dễ hiểu, và ông bảo vệ nó với sự khiêm nhường. Tuy nhiên, tôi nghĩ ông đã phạm sai lầm khi chiếm hữu cái tên "Singularity" và gán cho nó một nghĩa mới đầy lạc quan. Nó lạc quan đến nỗi khiến tôi, cũng như Vinge, cảm thấy cái nghĩa đó đáng sợ, được lấp đầy bằng những hình ảnh và ý tưởng hấp dẫn, che giấu những nguy hiểm. Ông đã dán cái nhãn mới, dìm yếu tố tai họa của AI xuống và quá thổi phồng lời hứa hẹn. Khởi nguồn từ một đề xướng công nghệ, Kurzweil đã tạo ra một trào lưu văn hóa mang âm hưởng tôn giáo mạnh mẽ. Tôi nghĩ rằng pha trộn sự biến đổi công nghệ với tôn giáo là một sai lầm lớn.

> Hãy tưởng tượng một thế giới khi mà khác biệt giữa người và máy mờ đi, khi ranh giới giữa nhân tính và công nghệ nhạt nhòa, khi linh

hồn và con chip silicon hợp nhất… Trong bàn tay đầy cảm hứng [của Kurzweil], cuộc sống ở thiên niên kỷ mới dường như không còn khiến người ta nản chí. Thay vào đó, thế kỷ 21 của Kurzweil hứa hẹn là một thời đại trong đó sự kết hợp giữa cảm tính con người với trí tuệ nhân tạo sẽ thay đổi và cải thiện về cơ bản cách chúng ta sống.

— *The Age of Spiritual Machines*
(bản bỏ túi, 1999)

Kurzweil không chỉ là cha đẻ của vấn đề Singularity, một người tranh luận lịch lãm và không bao giờ nao núng, mà còn là một người khởi xướng phong trào không biết mệt mỏi, thậm chí có phần máy móc. Ông là lãnh tụ của nhiều nam thanh và một số nữ tú, những người sống để chờ Singularity. Những người theo thuyết Singularity thường khoảng từ 20 đến hơn 30 tuổi, chủ yếu là đàn ông và không có con. Phần lớn họ là những anh chàng da trắng thông minh, những người đã nghe được tiếng gọi của Singularity. Nhiều người đã đáp lại bằng cách từ bỏ sự nghiệp, thứ vốn có thể làm bố mẹ họ tự hào, để sống như tu sĩ, tận tâm với các vấn đề Singularity. Đa phần họ tự học, một phần có lẽ vì không có chương trình đại học nào có chuyên ngành tổng hợp về khoa học máy tính, đạo đức, kỹ thuật sinh học, khoa học thần kinh, tâm lý học và triết học, nói ngắn gọn là nghiên cứu Singularity. (Kurzweil đồng sáng lập Đại học Singularity, một ngôi trường không cấp bằng và không được công nhận. Nhưng nó hứa hẹn "một sự hiểu biết rộng, đa chiều về những ý tưởng và vấn đề lớn nhất trong những công nghệ sẽ thay đổi tương lai.") Dù sao thì nhiều người theo thuyết Singularity cũng quá thông minh và tự lập nên không muốn đi theo con đường học hành truyền thống. Còn nhiều người khác thì là những kẻ dở hơi điên rồ mà rất ít trường cao đẳng hoặc đại học chịu nhận.

Một số người theo thuyết Singularity coi sự duy lý là giáo lý chủ yếu trong tín ngưỡng của mình. Họ tin rằng khả năng duy lý và logic tốt hơn, đặc biệt ở những người ra quyết định trong tương lai, sẽ giảm thiểu xác suất chúng ta tự tiêu diệt mình bằng AI. Họ lập luận, bộ não chúng ta đầy các thành kiến và suy nghiệm (heuristics) kỳ lạ, thứ vốn phục vụ tốt cho chúng ta trong quá trình tiến hóa, nhưng sẽ khiến chúng ta vướng vào rắc rối khi phải đối mặt với những lựa chọn và rủi ro đầy phức tạp trong thế giới hiện đại. Tiêu điểm chính của họ không phải là một Singularity tiêu cực và tai họa, mà là một thứ tích cực, sung sướng. Ở đó, chúng ta có thể tận dụng các công nghệ kéo dài tuổi thọ để sống thật lâu, có lẽ là ở dạng máy thay vì ở dạng sinh học. Nói cách khác, hãy tự thanh tẩy bạn khỏi những ý nghĩ sai lầm, và bạn có thể tìm thấy sự giải thoát khỏi thế giới nhục thể, và khám phá sự trường sinh bất lão.

Không có gì đáng ngạc nhiên khi Singularity thường được gọi là Khải huyền của Những kẻ cuồng công nghệ – trong vai trò của một phong trào, nó có những dấu hiệu đặc trưng của một tôn giáo ngày tận thế, bao gồm các nghi thức rửa tội, khuynh hướng xa rời thân xác tạm bợ của con người, chờ đợi cuộc sống vĩnh hằng, có một giáo chủ rõ ràng và (tương đối) lôi cuốn. Tôi hết lòng tán thành ý tưởng của những người theo thuyết Singularity, rằng AI là thứ quan trọng nhất hiện nay mà chúng ta nên quan tâm. Nhưng khi câu chuyện lái sang sự bất tử, tôi không có hứng thú. Những giấc mơ về cuộc sống vĩnh cửu có sức mạnh bóp méo sự thật một cách quá đáng. Quá nhiều người theo thuyết Singularity tin rằng sự hợp lưu của các dòng công nghệ hiện đang tăng tốc sẽ không sinh ra những loại thảm họa mà từng dòng công nghệ đơn lẻ có thể tạo ra, hoặc những thảm họa kết hợp mà chúng ta đã tiên đoán, mà ngược lại nó sẽ làm một thứ

khác biệt 180 độ. NÓ sẽ giải thoát loài người khỏi thứ họ lo sợ nhất: cái chết.

Làm thế nào mà bạn có thể đánh giá chính xác một số công cụ, và liệu sự phát triển của chúng có nên được điều chỉnh không, và bằng cách nào, khi mà bạn tin rằng chính những công cụ ấy sẽ cho phép bạn trường sinh bất lão? Thậm chí ngay cả những người duy lý bậc nhất trên thế giới cũng không có cái năng lực diệu kỳ để đánh giá tôn giáo của họ với một cái đầu lạnh. Và như học giả William Grassie◎ đã lập luận, khi bạn đang hỏi những câu hỏi về sự hóa thân, về số ít người được lựa chọn, và về cuộc sống vĩnh hằng, bạn đang nói về cái gì nếu không phải là tôn giáo?

> Liệu Singularity có dẫn tới việc các máy móc có tâm hồn sẽ thay thế con người? Hay sẽ dẫn tới việc con người hóa thân thành các siêu nhân bất tử sống trong một thiên đàng duy lý và khoái lạc? Con đường dẫn tới Singularity có phải đi qua một thời kỳ đau khổ? Liệu sẽ có một số ít người được bầu ra biết về bí mật của Singularity, những người tiên phong, và có lẽ sẽ là những tàn dư cuối cùng của nhân loại, đến được Miền đất hứa? Các chủ đề đậm tính tôn giáo như vậy đều hiện diện trong những bài hùng biện và lập luận duy lý của những người theo thuyết Singularity, thậm chí ngay cả khi những cách giải thích kiểu thuyết tiền và hậu thiên hy niên◎ không được phát triển một cách thích hợp, y như trong trường hợp của các phong trào Chúa cứu thế tiền khoa học.[6]

Không giống như quan điểm của Good và Vinge về tương lai tăng tốc, Singularity của Kurzweil được mang tới không chỉ nhờ trí tuệ nhân tạo, mà từ ba công nghệ tăng tiến tới điểm giao hòa – công nghệ gen, công nghệ

nano và công nghệ robot, một thuật ngữ rộng ông dùng để chỉ AI. Cũng không giống với Good và Vinge, Kurzweil đã đi tới một lý thuyết hợp nhất về tiến hóa công nghệ, thứ giống như bất cứ lý thuyết khoa học đứng đắn nào, cố gắng giải thích hiện tượng quan sát được và đưa ra những tiên đoán về các hiện tượng sẽ có trong tương lai. Nó được gọi là Định luật Hồi quy Tăng tốc, hay LOAR (Law of Accelerating Returns).

Đầu tiên Kurzweil đề xuất rằng các quá trình tiến hóa đi theo một đường cong mượt mà theo dạng hàm mũ, và sự phát triển của công nghệ chính là một trong các quá trình tiến hóa đó. Cũng giống như sự tiến hóa sinh học, công nghệ cho phép tiến hóa ra một năng lực, rồi dùng năng lực đó để tiến hóa tiếp lên nấc tiếp theo. Ví dụ như ở con người, bộ não lớn và ngón cái đối diện với lòng bàn tay cho phép chế tạo ra các công cụ và tạo ra sức nắm chặt để tay có thể sử dụng chúng một cách hiệu quả. Trong công nghệ, máy in góp phần vào công nghệ đóng sách, khiến dân chúng biết đọc biết viết, dẫn tới sự ra đời của các trường đại học và nhiều phát minh hơn nữa. Động cơ hơi nước tạo nên sức bật cho cách mạng công nghiệp và từ đó ngày càng thêm nhiều phát minh.

Công nghệ khởi đầu chậm do cách thức tự xây dựng trên cái nền là chính bản thân nó, nhưng sau đó đường cong tăng trưởng của nó dốc lên cho đến khi nó bắn thẳng lên gần như dốc đứng. Theo những biểu đồ và bảng biểu đã trở thành thương hiệu của Kurzweil, chúng ta đang bước vào thời kỳ quan trọng nhất của sự tiến hóa công nghệ, giai đoạn bắt đầu dốc đứng, cái "đầu gối của đường cong hàm mũ." Kể từ đây, nó chỉ có đi lên.

Kurzweil đã phát triển Định luật Hồi quy Tăng tốc của ông để mô tả sự tiến hóa của bất cứ quá trình nào mà trong đó có sự tiến hóa của các hình mẫu thông tin. Ông áp dụng LOAR cho sinh học, vốn ưu tiên sự gia tăng

tính trật tự của các phân tử, nhưng nó có tính thuyết phục hơn khi được sử dụng để dự đoán tốc độ thay đổi của công nghệ thông tin, bao gồm máy tính, máy ảnh kỹ thuật số, Internet, điện toán đám mây, thiết bị chẩn đoán và điều trị trong y tế, và nhiều nữa – bất cứ công nghệ nào có liên quan tới việc lưu trữ và thu hồi thông tin.

Như Kurzweil đã ghi chú, LOAR về bản chất là một lý thuyết kinh tế. Những Hồi quy Tăng tốc được các phát minh, sự cạnh tranh, thị phần hỗ trợ – những đặc trưng của thị trường và các công ty sản xuất. Trong thị trường máy tính, hiệu ứng này được biểu thị bằng Định luật Moore, một lý thuyết kinh tế khác đội lốt một lý thuyết công nghệ, được nhận ra lần đầu tiên vào năm 1965 bởi nhà đồng sáng lập Intel là Gordon Moore.

Định luật Moore nói rằng số lượng transistor có thể gắn lên một vi mạch để chế tạo ra một chip vi xử lý sẽ tăng gấp đôi sau mỗi chu kỳ 18 tháng. Một transistor là một công tắc bật/tắt mà cũng có thể khuếch đại điện áp. Càng nhiều transistor nghĩa là tốc độ xử lý càng lớn, và máy tính chạy càng nhanh. Định luật Moore hàm nghĩa máy tính sẽ trở nên nhỏ hơn, mạnh hơn và rẻ hơn với một tốc độ ổn định. Chuyện này không xảy ra nhờ Định luật Moore là một định luật của thế giới vật lý, kiểu như định luật hấp dẫn, hoặc Định luật Thứ hai về Nhiệt động học. Nó xảy ra vì người tiêu dùng và thị trường thương nghiệp luôn thúc đẩy người sản xuất chip máy tính phải cạnh tranh và góp phần tạo ra những máy tính, điện thoại thông minh, máy ảnh, máy in, pin năng lượng mặt trời, và sắp tới là máy in 3D, ngày một nhỏ hơn, nhanh hơn và rẻ hơn. Người sản xuất chip đang sử dụng các công nghệ và kỹ thuật của quá khứ. Vào năm 1971, 2.300 transistor có thể được gắn lên một con chip. 40 năm sau, hay là sau hai lần 20, con số đó là

2.600.000.000.[7] Và với sự tăng tốc như vậy, hơn hai triệu transistor nói trên có thể được gắn vào dấu chấm ở cuối câu này.

Sau đây là một ví dụ ngoạn mục cho lập luận trên. Jack Dongarra, một nhà nghiên cứu ở Phòng nghiên cứu quốc gia Oak Ridge thuộc bang Tennessee, thành viên của một đội chuyên theo dõi tốc độ của các siêu máy tính, xác định rằng máy tính bảng bán chạy nhất của Apple, iPad 2, có tốc độ tương đương siêu máy tính Cray 2 vào khoảng năm 1985. Thực tế là với tốc độ 1,5 gigaflop (1 gigaflop bằng một *tỉ* thao tác tính toán, hay phép toán trên một giây), iPad 2 hẳn đã nằm trong danh sách 500 siêu máy tính chạy nhanh nhất thế giới hồi năm 1994.[8]

Vào năm 1994, ai mà có thể tưởng tượng rằng chưa đầy một thế hệ nữa, một siêu máy tính với kích cỡ nhỏ hơn một cuốn sách sẽ đủ rẻ để có thể cho không học sinh cấp 3, và hơn thế, nó sẽ kết nối vào kho kiến thức nhân loại mà *không cần* dây cáp? Chỉ có Kurzweil mới táo bạo như vậy, và dù ông không đưa ra lời xác nhận chính xác về các siêu máy tính, nhưng ông đã dự đoán đúng sự bùng nổ của Internet.[9]

Trong công nghệ thông tin, mỗi bước đột phá lại kéo theo bước đột phá kế tiếp đến gần hơn – đường cong chúng ta đã nói đến sẽ dốc thêm. Do vậy, khi nghĩ về iPad 2, câu hỏi đặt ra không phải là chúng ta sẽ chờ đợi điều gì sau 15 nam *tiếp theo*. Thay vào đó, hãy chú ý đến những gì sẽ xảy ra trong một phần nhỏ của thời lượng đó. Vào khoảng năm 2020, Kurzweil ước đoán rằng chúng ta sẽ sở hữu các laptop với sức mạnh xử lý thô tương đương bộ não người, cho dù chúng chưa có sự thông minh.

Chúng ta hãy cùng xem Định luật Moore áp dụng cho sự bùng nổ trí thông minh thế nào. Nếu chúng ta giả định rằng AGI có thể đạt được, Định

luật Moore sẽ hàm ý rằng ngay cả các vòng lặp tự cải tiến của sự bùng nổ trí thông minh cũng không nhất thiết cần phải có để đạt được ASI, hay siêu trí tuệ nhân tạo. Đó là vì một khi bạn đã đạt được AGI, hai năm sau các máy móc với trí thông minh cấp độ con người sẽ tăng tốc gấp đôi. Và sau hai năm nữa, lại tiếp tục tăng gấp đôi. Trong khi đó, trí thông minh trung bình của con người vẫn y như cũ. AGI sẽ nhanh chóng vượt xa chúng ta.

Nếu trí thông minh đó bắt đầu tham gia quá trình tự cải tiến bản thân nó, thì điều gì sẽ xảy ra? Eliezer Yudkowsky mô tả việc sau khi có AGI, tốc độ tăng trưởng công nghệ sẽ vuột khỏi tầm điều khiển của chúng ta nhanh đến thế nào.

Nếu tốc độ máy tính tăng gấp đôi mỗi hai năm, điều gì xảy ra nếu một AI được cài trên máy tính đang thực hiện các nghiên cứu?

> Tốc độ máy tính tăng gấp đôi mỗi hai năm.
>
> Tốc độ máy tính tăng gấp đôi mỗi hai năm làm việc.
>
> Tốc độ máy tính tăng gấp đôi mỗi hai năm làm việc theo hệ quy chiếu hiện tại.
>
> Hai năm sau khi Trí tuệ nhân tạo đạt cấp độ con người, tốc độ của chúng nhân đôi. Một năm sau, tốc độ của chúng nhân đôi tiếp. 6 tháng – 3 tháng – 1,5 tháng… Singularity.[10]

Một số người phản đối rằng Định luật Moore sẽ dừng lại trước năm 2020, khi chuyện cho thêm transistor vào các vi mạch trở nên bất khả thi về mặt vật lý. Một số khác nghĩ rằng Định luật Moore sẽ tạo điều kiện cho sự nhân đôi *nhanh hơn* khi các bộ vi xử lý được áp dụng công nghệ mới,

dùng những bộ phận nhỏ hơn để thực hiện việc tính toán, chẳng hạn các phân tử, các hạt photon ánh sáng, thậm chí cả DNA. Các con chip xử lý 3D, được Đại học Bách khoa Federale de Lausanne của Thụy Sĩ (EPFL) phát triển, có thể là thứ đầu tiên thắng được Định luật Moore. Dù chưa đi vào sản xuất, nhưng các con chip của EPFL được sắp theo phương thẳng đứng thay vì trên một mặt phẳng như trước đây, và sẽ nhanh hơn, có hiệu suất cao hơn các con chip truyền thống, cũng như dùng được trong công nghệ xử lý song song.[11] Và ngay cả công ty mà Gordon Moore đồng sáng lập cũng đã nối dài Định luật Moore với việc sáng chế ra *transistor* 3D đầu tiên. Như đã nói, transistor là các công tắc điện tử. Transistor truyền thống hoạt động bằng cách điều chỉnh dòng điện chạy trong không gian hai chiều. Transistor Tri-Gate mới của Intel điều khiển dòng điện chạy trong không gian ba chiều, với tốc độ tăng 30% và tiết kiệm 50% năng lượng. Một tỉ transistor Tri-Gate sẽ có mặt trong mỗi con chip thế hệ mới của Intel[12]

Vì việc gắn các transistor lên silicon được dùng nhiều trong công nghệ thông tin, từ máy ảnh đến các cảm biến y tế, nên Định luật Moore cũng áp dụng cho chúng. Nhưng Moore đã lập ra lý thuyết về các vi mạch chứ không phải về nhiều lĩnh vực được kết nối trong công nghệ thông tin, vốn bao gồm cả cách thức sản xuất và sản phẩm. Do đó, định luật tổng quát hơn của Kurzweil, Định luật Hồi quy Tăng tốc, là một lựa chọn thích hợp hơn. Và nhiều công nghệ đang *trở thành* công nghệ thông tin, như máy tính, và thậm chí robot, chúng tăng trưởng và ngày càng có quan hệ mật thiết với mọi khía cạnh của thiết kế sản phẩm, sản xuất và bán hàng. Chú ý là việc sản xuất điện thoại thông minh – không chỉ con chip xử lý của nó – đã hưởng lợi từ cuộc cách mạng kỹ thuật số. Mới chỉ sáu năm kể từ ngày chiếc iPhone đầu tiên ra đời, Apple đã cho ra *sáu* phiên bản. Apple đã tăng

tốc độ của chúng lên gấp đôi, và với hầu hết người dùng thì giá thành đã giảm hơn một nửa. Đó là vì tốc độ phần cứng trong các thành phần của máy đã được nhân đôi như thường lệ. Và ngay trong từng công đoạn của dây chuyền sản xuất, sự nhân đôi đó cũng diễn ra.

Các hiệu ứng được LOAR dự đoán vươn xa hơn cả ngành công nghiệp máy tính và điện thoại thông minh. Gần đây nhà đồng sáng lập Google là Larry Page đã gặp Kurzweil để thảo luận về tình trạng nóng lên toàn cầu, và họ đã chia tay với một tâm trạng lạc quan. Sau 20 năm nữa, họ khẳng định, công nghệ nano sẽ khiến năng lượng mặt trời trở nên kinh tế hơn dầu mỏ hoặc than đá. Ngành công nghiệp này sẽ cung cấp cho thế giới 100% số năng lượng mà nó cần. Dù năng lượng mặt trời chỉ đủ dùng cho 0,5% nhu cầu hiện nay của thế giới, nhưng họ cho rằng con số đó sẽ nhân đôi mỗi hai năm như điều đã xảy ra suốt 20 năm qua. Vậy là hai năm nữa năng lượng mặt trời sẽ chiếm 1% sản lượng năng lượng toàn cầu, sau bốn năm nữa là 2%, và sau 16 năm nữa sẽ nhân đôi tám lần bằng 256% nhu cầu năng lượng của thế giới. Ngay cả khi tính đến sự tăng dân số và nhu cầu năng lượng lớn hơn trong hai thập niên sau đấy, từng đó năng lượng mặt trời vẫn đủ và còn thừa ra một ít. Và như thế, theo Kurzweil và Page, tình trạng nóng lên toàn cầu sẽ được giải quyết.[13]

Và với cái chết cũng sẽ như thế. Theo Kurzweil, chuyện kéo dài sự sống mãi mãi hầu như đã ở trong tầm tay.

"Chúng ta hiện giờ đã có những phương tiện thực sự để hiểu rõ phần mềm của cuộc sống và tái lập trình nó; chúng ta có thể dừng sự hoạt động của các gen mà không phải can thiệp vào, có thể thêm vào các gen mới và cả những bộ phận cơ thể mới với liệu pháp tế bào gốc," Kurzweil nói. "Điểm mấu chốt ở đây, đó là y học hiện nay là một dạng công nghệ thông

tin – sức mạnh của nó sẽ nhân đôi mỗi năm. Những công nghệ này sẽ mạnh hơn một triệu lần sau 20 năm, với cùng một giá thành."[14]

Kurzweil tin rằng con đường ngắn nhất dẫn tới AGI là kỹ nghệ đảo ngược bộ não – quét nó vô số lần để thu được một bộ tổng hợp các vi mạch của não. Được thay thế bằng các giải thuật hoặc các mạng phần cứng, những vi mạch này sẽ được khởi động trên một máy tính như là một bộ não tổng hợp duy nhất, và được dạy mọi thứ nó cần biết. Nhiều tổ chức đang làm việc trong các dự án đi theo con đường tới AGI này. Chúng ta sẽ thảo luận một số cách tiếp cận và các chướng ngại sắp tới.

• • •

Tốc độ tiến hóa của phần cứng cần thiết để chạy một bộ não ảo đòi hỏi một cái nhìn sâu hơn. Chúng ta hãy bắt đầu với bộ não người và sau đó hướng tới những máy tính có thể giả lập chúng, Kurzweil viết rằng bộ não có khoảng 100 tỉ neuron, mỗi neuron kết nối với khoảng 1.000 neuron khác.[◎] Như vậy là khoảng 100 triệu triệu kết nối Mỗi kết nối có khả năng thực hiện khoảng 200 tính toán mỗi giây (các vi mạch nhanh hơn thế ít nhất là 10 triệu lần). Kurzweil nhân các kết nối bên trong neuron của bộ não với tốc độ tính toán từng cái và nhận được kết quả là 20 triệu tỉ tính toán trên một giây, hoặc 20.000.000.000.000.000.[16]

Danh hiệu siêu máy tính nhanh nhất thay đổi gần như hằng tháng, nhưng ngay lúc này thì máy tính Sequoia của Bộ Năng lượng đang chiếm ngôi với hơn 16 petaflop. Đó là 16 triệu tỉ tính toán trên một giây, hoặc là khoảng 80% tốc độ của bộ não người, như Kurzweil đã tính vào năm 2000.[17] Tuy nhiên trong cuốn *The Singularity is Near*, vào năm 2005, Kurzweil đã giảm ước lượng tốc độ bộ não từ 20 xuống còn 16 petaflop, và

dự đoán rằng một siêu máy tính sẽ đạt được tốc độ ấy vào năm 2013.[18] Sequoia đã làm được sớm hơn một năm.

Có thật là chúng ta đã tới gần giới hạn của bộ não? Những con số này có thể dễ gây nhầm lẫn. Bộ não là những vi xử lý song song và làm cực tốt một số việc, trong khi máy tính hoạt động tuần tự và làm tốt những việc khác. Bộ não thì chậm và làm việc với những xung mạch của các neuron. Máy tính có thể xử lý công việc nhanh hơn và lâu hơn, thậm chí lâu vô hạn.

Tuy thế, bộ não người vẫn là ví dụ duy nhất của chúng ta về trí thông minh cao cấp. Nếu kỹ thuật "brute force" muốn cạnh tranh với nó, thì máy tính phải thực hiện được những chiến công ấn tượng về nhận thức. Hãy xem một số mẫu siêu máy tính thường gặp trong các hệ thống phức tạp hiện nay: các hệ thống thời tiết, vụ nổ hạt nhân 3D và giả lập phân tử dùng trong sản xuất. Bộ não người liệu có bao hàm một sự phức tạp giống thế, hoặc hơn thế không? Theo mọi chỉ báo thì chúng tương đương nhau.

Có lẽ, như Kurzweil nói, chinh phục bộ não người chỉ là chuyện nay mai, và 30 năm sắp tới của khoa học máy tính sẽ tương đương 140 năm, tính theo tốc độ tiến triển hiện nay.[19] Phải nói thêm rằng *chế tạo* AGI cũng là một công nghệ thông tin. Với tốc độ máy tính tăng theo hàm mũ, các nhà nghiên cứu AI có thể tăng tốc công việc của họ. Nó có nghĩa là viết nhiều giải thuật phức tạp hơn, nhiều giải thuật thiên về sức mạnh hơn, giải quyết các vấn đề điện toán khó hơn và tao ra nhiều thử nghiệm hơn.[20] Các máy tính nhanh hơn góp phần tạo ra một ngành công nghiệp AI mạnh hơn, đến lượt mình sẽ cung cấp thêm nhiều nhà nghiên cứu máy tinh và các công cụ hữu dụng để đạt tới AGI.[21]

Kurzweil viết rằng sau khi các nhà nghiên cứu chế tạo ra một máy tính có khả năng vượt qua được bài kiểm tra Turing vào năm 2029, mọi thứ thậm chí sẽ còn tăng tốc mạnh hơn.[22] Nhưng ông không tiên liệu về một vụ bùng phát Singularity cho đến tận 16 năm sau, 2045. Sau đó tốc độ tiến bộ công nghệ sẽ vượt quá khả năng điều khiển của bộ não chúng ta. Do đó, ông lập luận, chúng ta cần phải tăng cường bộ não để theo kịp. Điều đó có nghĩa là đưa những công nghệ trợ giúp não trực tiếp vào các mạch neuron của chúng ta, giống như cách mà hiện nay các thiết bị cấy ở lỗ tai kết nối trực tiếp đến các dây thần kinh thính giác trợ giúp cho người khiếm thính.[23] Chúng ta sẽ làm cho những kết nối neuron chậm chạp đó chạy tốt hơn, nghĩ nhanh hơn, sâu sắc hơn và nhớ được nhiều hơn. Chúng ta sẽ truy cập đến tất cả kiến thức của nhân loại, và tương tự như máy tính, sẽ có khả năng chia sẻ những suy nghĩ và kinh nghiệm của mình tới người khác một cách tức thời, trong khi trải nghiệm những suy nghĩ của họ. Sau cùng, công nghệ sẽ giúp chúng ta nâng cấp bộ não với một bộ nhớ trung gian bền hơn mô não, hoặc chúng ta sẽ tải ý thức của mình lên máy tính, trong khi vẫn bảo tồn những phẩm chất khiến chúng ta là *chúng ta*.

Tất nhiên, viễn cảnh tương lai này giả định một điều là *bạn* ở trong bạn, bản thân bạn, có thể di chuyển, và đó là một giả định rất lớn. Nhưng đối với Kurzweil, đây là con đường đến với sự bất tử, là một biển cả kiến thức và kinh nghiệm vượt quá những gì con người hiện tại có thể nắm bắt. Sự tăng cường trí thông minh sẽ xảy ra thật từ tốn khiến ít người gạt bỏ nó. Nhưng "từ tốn" nghĩa là vào năm 2045, vậy là chỉ 30 năm nữa, với những thay đổi chủ yếu, sâu sắc diễn ra vào năm năm cuối cùng, hoặc gần như thế. Như vậy có tuần tự không? Tôi nghĩ là không.

Như chúng ta đã lưu ý ở trên, Apple đã cho ra sáu phiên bản iPhone trong sáu năm. Theo Định luật Moore, trong khoảng thời gian đó phần cứng của họ đã tăng trưởng đủ để có được hai hoặc nhiều hơn thế số lần nhân đôi, nhưng nó chỉ tăng một lần. Tại sao? Đó là vì thời gian chậm trễ sinh ra trong quá trình phát triển, thử nghiệm và sản xuất các linh kiện của iPhone, bao gồm bộ vi xử lý, camera, bộ nhớ, ổ cứng, màn hình… và sau đó là việc tiếp thị và bán hàng.

Liệu khoảng thời gian trễ từ tiếp thị đến bán hàng này có bao giờ mất đi? Có lẽ đến một ngày nào đó, phần cứng cũng sẽ tự cập nhật như là phần mềm hiện nay. Nhưng chuyện đó sẽ không xảy ra cho đến khi khoa học làm chủ được công nghệ nano hoặc kỹ nghệ in 3D trở nên phổ cập. Và khi nâng cấp những bộ phận của bộ não chúng ta, thay vì cập nhật phần mềm Microsoft Office hoặc đi mua vài con chip hay RAM, đó sẽ là một chuỗi hành động tinh xảo hơn bất cứ những gì chúng ta từng trải nghiệm, ít nhất là trong lần đầu tiên.

Vậy mà Kurzweil tuyên bố rằng trong thế kỷ này, chúng ta sẽ trải nghiệm 200.000 năm tiến bộ công nghệ trong 100 năm lịch sử. Liệu chúng ta có thể chịu được nhiều sự thay đổi diễn ra nhanh chóng đến vậy?

Nicholas Carr, tác giả cuốn *The Shallows* (Những kẻ hời hợt), lập luận rằng điện thoại thông minh và máy tính đang làm giảm chất lượng tư duy của chúng ta và thay đổi hình thái của bộ não người. Trong cuốn *Virtually You* (Bạn trong Internet), nhà tâm thần học Elias Ahoujaoude cảnh báo rằng các mạng xã hội và trò chơi nhập vai đã cổ vũ cho một lượng lớn những thứ bệnh hoạn, bao gồm thói ái kỷ và tính vị kỷ quá mức. Sự đắm chìm vào công nghệ làm yếu đi cá tính và bản ngã – là nhận định của Jaron Lanier, lập trình viên, người *tiên phong* trong công nghệ thực tế ảo, tác giả của

cuốn *You Are Not a Gadget: A Manifesto* (Bạn không phải là một thứ phụ tùng: Lời tuyên ngôn). Các chuyên gia này cảnh báo những hậu quả bất lợi đến từ những máy tính *bên ngoài* cơ thể chúng ta. Và Kurzweil đề xuất rằng những thứ tốt đẹp sẽ chỉ đến từ những máy tính *bên trong* cơ thể chúng ta. Tôi nghĩ thật không hợp lý khi chờ đợi điều vốn từng diễn ra trong hàng trăm ngàn năm tiến hóa sẽ thay đổi chỉ trong 30 năm, và chúng ta có thể được tái lập trình để yêu thích một sự tồn tại, thứ vô cùng khác biệt với cuộc sống mà chúng ta đã mất bao thế hệ để thích nghi.

Nhiều khả năng con người sẽ quyết định một tốc độ thay đổi mà họ có thể chế ngự và kiểm soát được. Mỗi người có thể chọn một cách khác nhau, và nhiều người sẽ chọn cùng một tốc độ cũng tương tự như việc họ cùng thích một kiểu quần áo, ô tô hoặc máy tính. Chúng ta đã biết rằng Định luật Moore và LOAR mang tính kinh tế thay vì tính tất định. Nếu có đủ người với đủ tài nguyên muốn tăng tốc bộ não của họ một cách nhân tạo, họ sẽ tạo ra một lượng cầu *nhất định*. Tuy nhiên, tôi nghĩ Kurzweil đã đánh giá quá cao ham muốn được nghĩ nhanh hơn và sống lâu hơn của các cá nhân trong tương lai. Dù sao đi nữa, tôi không nghĩ là một Singularity tràn đầy niềm vui, như ông đã định nghĩa về nó, sẽ trở thành hiện thực. AI được phát triển mà không có đủ độ an toàn sẽ làm hỏng nó.

Công cuộc chế tạo AGI là bất khả kháng và có lẽ là bất khả trị. Và vì thứ động lực nhân đôi đã được biểu thị bằng LOAR, AGI sẽ chiếm lấy sân khấu thế giới (ý tôi là *chiếm)* sớm hơn nhiều so với những gì ta nghĩ.

10

Những Người Theo Thuyết Singularity

> Khác với trí tuệ của chúng ta, cứ 18 tháng máy tính lại nhân đôi hiệu suất. Vậy chuyện chúng sẽ trở nên thông minh và chiếm lấy thế giới là mối nguy có thật.[1]
>
> — **Stephen Hawking**
> nhà Vật lý

> Trong vòng 30 năm, chúng ta sẽ có những phương tiện công nghệ để tạo ra trí thông minh siêu nhân. Ngay sau đó, kỷ nguyên con người sẽ kết thúc. Có thể tránh được tiến trình như vậy không? Nếu không thể tránh khỏi, liệu con người có cách nào lèo lái nó để t`ồn tại?[2]
>
> — **Vernor Vinge**
> tác gia, giáo sư, nhà Khoa học Máy tính

Kể từ năm 2005, Viện nghiên cứu Trí thông minh Máy tính, tên cũ là Viện Singularity v`ềTrí tuệ Nhân tạo, đ`âu tổ chức Hội nghị thượng đỉnh Singularity thường niên. Trong hai ngày một loạt các diễn giả sẽ diễn

thuyết cho khoảng 1.000 thành viên về bối cảnh lớn của Singularity – tác động của nó đến việc làm và nền kinh tế, sức khỏe và sự trường thọ, cùng những hệ quả đạo đức của nó. Diễn giả của Hội nghị năm 2011 tại thành phố New York bao gồm các huyền thoại khoa học như Stephen Wolfram của Mathematica, tỉ phú dot-com Peter Thiel, người vẫn thường tài trợ để những người trẻ đam mê công nghệ có thể không học đại học và khởi nghiệp, David Ferrucci của IBM, chủ nhiệm dự án DeepQA/Watson. Eliezer Yudkowsky luôn diễn thuyết, và thường sẽ có một hoặc hai nhà đạo đức học cũng như những người phát ngôn của các cộng đồng trường sinh◎ và siêu hóa.◎ Những người phái trường sinh khảo sát các công nghệ và liệu pháp cho phép con người sống mãi mãi. Những người phái siêu hóa suy nghĩ về phương thức sử dụng phần cứng và mỹ phẩm để tăng cường năng lực cũng như vẻ đẹp của con người, và… những cơ hội để sống mãi. Đứng trên tất cả những nhóm đó là Người khổng lồ của Singularity, đồng sáng lập Hội nghị thượng đỉnh Singularity, và là ngôi sao của mọi hội thảo, Ray Kurzweil.

Chủ đề của Hội nghị năm 2011 là máy tính DeepQA Watson của IBM, và Kurzweil đã trình bày một bài diễn thuyết tẻ ngắt về lịch sử của các chatbot◎ và các hệ thống Hỏi-Đáp, có tên là "từ Eliza đến Watson." Nhưng vào giữa cuộc nói chuyện, ông phấn chấn hẳn lên khi phản bác lại một bài tiểu luận vụng về tấn công giả thuyết Singularity của ông, thứ một phần do người đồng sáng lập Microsoft là Paul Allen chắp bút.

Kurzweil trông không thực sự khỏe – ông gầy, hơi ngập ngừng, trầm lắng hơn mọi khi. Ông không phải là kiểu diễn giả có khả năng nuốt trọn sân khấu hoặc thổi bùng nhiệt huyết. Ngược lại, ông có một tác phong nhẹ nhàng, hơi máy móc, loại tác phong dường như dành cho các cuộc thương

lượng con tin hoặc đọc truyện đêm khuya. Nhưng nó lại rất hợp với những ý tưởng cách mạng ngẫu hứng mà ông vẫn nói đến. Trong thời đại của những tỉ phú dot-com mặc quần jean đi diễn thuyết, Kurzweil mặc loại quần dài màu nâu của thế hệ trước, cùng với giày lười có núm tua, áo khoác thể thao và đeo kính. Người ông không to cũng không nhỏ, nhưng gần đây bắt đầu có vẻ già, đặc biệt là khi so sánh với một Kurzweil đầy sinh lực trong ký ức của tôi. Ông mới chỉ 52 hoặc xấp xỉ như thế trong lần gần nhất tôi phỏng vấn ông, khi chưa bắt đầu ăn kiêng ít calo ngặt nghèo, hiện là một phần trong kế hoạch làm chậm quá trình lão hóa của ông. Với một chế độ được nghiên cứu kỹ lưỡng bao gồm ăn kiêng, tập thể dục, và uống các loại vitamin, Kurzweil muốn đẩy lùi tử thần cho đến khi công nghệ tìm được thứ thần dược mà ông chắc chắn nó sẽ xuất hiện.

"Tôi là một kẻ lạc quan. Tôi cần phải như thế, trong vai trò một nhà phát minh."

Sau bài diễn thuyết, Kurzweil và tôi ngồi cạnh nhau trên những cái ghế kim loại trong một phòng thay đồ nhỏ ở tầng trên sân khấu. Một đoàn làm phim tài liệu đang chờ ở ngoài để phỏng vấn ông khi tôi đã xong. Chỉ một thập niên trước, khi ông còn là một nhà phát minh và tác giả chưa mấy tiếng tăm, tôi đã chiếm dụng ông trong ba giờ đồng hồ thú vị cùng với đội làm phim của mình; còn giờ ông là một nhân vật nổi tiếng mà tôi chỉ có thể ở cạnh chừng nào còn giữ được cánh cửa ra vào khép kín. Bản thân tôi cũng đã thay đổi – cuộc gặp đầu tiên đã khiến tôi choáng ngợp trước ý tưởng đưa bộ não của mình vào máy tính, như đã được mô tả trong *Spiritual Machines*. Những câu hỏi của tôi ngây thơ trong veo như bong bóng rượu champagne. Giờ thì tôi đã hoài nghi hơn, và hiểu biết nhiều hơn

về những mối nguy mà đã không còn khiến bậc thầy này thấy đáng quan tâm.

"Tôi đã thảo luận tương đối nhiều về những hiểm họa trong *The Singularity is Near,"* Kurzweil phản đối khi tôi hỏi liệu ông có thổi phồng những triển vọng và dìm các mối nguy của Singularity xuống. "Chương 8 viết về sự gắn kết sâu sắc giữa triển vọng và hiểm họa của GNR (genetics, nanotechnology, robotics – công nghệ gen, công nghệ nano, công nghệ robot) và tôi đã mô tả khá chi tiết về những mặt trái của ba lĩnh vực công nghệ nói trên. Và mặt trái của công nghệ robot, thật ra là AI, thứ có ảnh hưởng sâu rộng nhất vì trí thông minh là hiện tượng quan trọng nhất trên thế giới. Về bản chất, không tồn tại một sự bảo vệ tuyệt đối nào trước AI mạnh cả."

Cuốn sách của Kurzweil có lưu ý về các hiểm họa của kỹ thuật di truyền và công nghệ nano, nhưng nó chỉ có vài trang yếu ớt về AI mạnh, cái tên cũ của AGI. Và trong chương đó ông cũng tranh luận rằng sự từ bỏ hay quay lưng lại với một số công nghệ vì chúng quá nguy hiểm, như Bill Joy và một số người khác đã chỉ ra, không chỉ là một ý tưởng tồi mà còn vô đạo đức. Tôi đồng ý rằng từ bỏ là bất khả thi. Nhưng còn vô đạo đức?

"Việc từ bỏ là vô đạo đức bởi nó sẽ cướp đi của chúng ta những lợi ích sâu sắc. Con người hiện vẫn có nhiều nỗi đau khổ mà họ có thể vượt qua, và vì thế việc đó là một sứ mệnh đạo đức. Thứ hai, sự từ bỏ đòi hỏi một hệ thống toàn trị để cấm công nghệ này. Thứ ba và cuối cùng, không thể từ bỏ được. Nó sẽ chỉ làm cho các công nghệ này phát triển trong thế giới ngầm, nơi mà những người thực hiện nó sẽ vô trách nhiệm và không bị giới hạn. Những nhà khoa học có trách nhiệm, có nhiệm vụ phát triển các hệ thống

phòng thủ, sẽ không có công cụ cần thiết để làm việc này. Và điều đó thật ra còn nguy hiểm hơn."

Kurzweil phê phán cái gọi là Nguyên tắc Phòng ngừa, một đề xuất của phong trào môi trường, cũng như sự từ bỏ, nó chỉ là một cách né tránh cuộc đối thoại này. Nhưng điều quan trọng là phải nhắc lại nguyên tắc đó để xem tại sao cách lý luận của nó không có trọng lượng. Nguyên tắc Phòng ngừa nói rằng: "Nếu những hậu quả của một hành động là không rõ ràng, nhưng được một số nhà khoa học dự đoán rằng nó ẩn chứa dù chỉ một rủi ro nhỏ có tác động tiêu cực sâu sắc, thì không làm việc đó sẽ tốt hơn là chấp nhận rủi ro."[3] Nguyên tắc này thường không được áp dụng hoặc nếu có thì cũng không chặt chẽ. Nó sẽ dừng bất cứ công nghệ nào được tuyên bố là nguy hiểm nếu "một số nhà khoa học" sợ nó, dù cho họ không liên quan gì đến dây chuyền những sự kiện dẫn tới kết cục mà họ lo ngại.

Nguyên tắc Phòng ngừa và sự từ bỏ không có triển vọng thành công khi áp dụng vào AGI. Cả hai phương sách này đều không thể thực hiện được, trừ phi có một tai nạn khủng khiếp trên con đường tới AGI làm chúng ta sợ cứng người. Những dự án AGI mạnh nhất của các công ty và chính phủ sẽ tìm kiếm ưu thế cạnh tranh của sự bảo mật – chúng ta đã thấy chuyện này ở các công ty ẩn danh rồi. Sẽ có rất ít quốc gia hoặc công ty chịu từ bỏ lợi thế này, kể cả nếu việc phát triển AGI được đặt ngoài vòng pháp luật. (Trên thực tế, Google có nguồn tài chính và ảnh hưởng ngang với một quốc gia/dân tộc hiện đại, do vậy để hình dung các nước khác sẽ làm gì, hãy theo dõi Google). Công nghệ cần thiết cho AGI hiện diện khắp nơi và dùng được vào nhiều mục đích và ngày càng thu hẹp về quy mô. Rất khó, nếu không nói là không thể, không chế sự phát triển của nó.

Nhưng liệu có vô đạo đức khi *không* phát triển AGI, như Kurzweil đã đề ra, lại là chuyện khác. Thứ nhất, các lợi ích của AGI là vô cùng to lớn, nhưng chỉ khi chúng ta còn sống mà hưởng chúng. Và đó là một chữ *nếu* cực to khi hệ thống đã tiến bộ đủ để khởi phát sự bùng nổ trí thông minh. Lập luận của Kurzweil là có vấn đề, khi cho rằng các lợi ích chưa chứng minh được sẽ lớn hơn các rủi ro chưa chứng minh được. Tôi đã nói với ông rằng tôi nghĩ sẽ là vô đạo đức nếu phát triển các công nghệ như AGI mà không đồng thời giáo dục càng nhiều người càng tốt về các hiểm họa của nó. Tôi nghĩ rằng các hiểm họa kinh khủng của AGI, hiện đã được nhiều nhà nghiên cứu tài năng và được tôn trọng công nhận, đã được xác nhận rõ ràng thay vì những lợi ích giả định về Singularity của ông – máu được làm sạch nhờ công nghệ nano, các bộ não nhanh và mạnh hơn, sự bất tử… Sự chắc chắn duy nhất về Singularity, đó là nó mô tả một thời kỳ mà trong đó, do sức mạnh của LOAR, chúng ta sẽ có những máy tính nhanh, thông minh, hoạt động trong mọi lĩnh vực của đời sống và cả trong cơ thể chúng ta. Sau đó trí thông minh máy móc xa lạ sẽ khiến trí thông minh vốn có của chúng ta phải trầm trồ sợ hãi. Liệu chúng ta có thích nó không lại là một chuyện khác.

Nếu bạn đọc Kurzweil kỹ càng, bạn sẽ thấy các lợi ích chủ yếu bắt nguồn từ IA, và IA là cần thiết để theo kịp với tốc độ thay đổi nhanh và sắc bén. Như đã lập luận, tôi nghĩ rằng nó sẽ đảo ngược niềm đam mê công nghệ của con người, thậm chí dẫn tới chán ghét.

Và điều đó thậm chí không phải là mối lo sợ chính của tôi, vì tôi không nghĩ chúng ta đến được thời điểm đó. Tòi nghĩ rằng những công cụ quá mạnh vượt quá tầm kiểm soát sẽ khiến chúng ta dừng lại giữa đường. Tôi nói với Kurzweil điều này và ông phản biện lại bằng vài lý luận cũ mèm –

những luận điểm lạc quan, nhân cách hóa giống hệt như thứ ông đã nói với tôi 10 năm trước.

"Việc có một thực thể rất thông minh và vì lý do nào đó lại muốn hủy diệt chúng ta là một kịch bản tiêu cực. Nhưng bạn phải hỏi tại sao lại có một thứ như thế? Trước hết tôi vẫn giữ quan điểm rằng chúng ta sẽ không đối đầu với máy móc, bởi những máy móc này không thuộc về một nền văn minh khác, mà là một phần của nền văn minh con người. Chúng là những công cụ mà chúng ta sử dụng, chúng là một thứ tay chân của chúng ta, và ngay cả khi chúng ta *trở thành* công cụ thì chúng vẫn chỉ là một thứ tiến hóa từ nền văn minh con người. Đó không phải là cuộc ngoại xâm của những máy móc đến từ sao Hỏa. Chúng ta sẽ không phải phân vân các giá trị của chúng là gì."

Như chúng ta đã thảo luận, việc giả định rằng AGI sẽ chỉ giống như chúng ta chính là sự gán ghép những giá trị của con người vào máy móc thông minh, thứ có trí tuệ và giá trị của nó theo một cách rất khác biệt với con người. Mặc dù người tạo ra chúng có ý định tốt, nhưng ở hầu hết, nếu không muốn nói là tất cả các AGI, chuyện hệ thống hoạt động thế nào sẽ là một điều quá mơ hồ *và* quá phức tạp với chúng ta để có thể hoàn toàn hiểu hoặc dự đoán. Xa lạ, không thể hiểu được, và cuối cùng là việc một số AGI sẽ được tạo ra với ý đồ giết người, vì đừng quên rằng ở Mỹ, các cơ quan quốc phòng là những nhà đầu tư tích cực nhất. Chúng ta nên thừa nhận rằng điều này cũng đúng ở các quốc gia khác.

Tôi chắc chắn rằng Kurzweil đã xem xét chuyện AGI không cần phải được thiết kế với mục đích làm hại con người để hủy diệt nhân loại, chỉ cần chúng phớt lờ con người là đủ. Như Steve Omohundro cảnh báo, thiếu đi sự lập trình cẩn trọng, AI cao cấp sẽ sở hữu các động lực và mục tiêu mà

chúng ta không có.[4] Như Eliezer Yudkowsky nói, nó có thể muốn sử dụng các phân tử của cơ thể con người vào mục đích khác.[5] Và như chúng ta đã thấy, AI thân thiện, thứ sẽ bảo đảm AGI đầu tiên và tất cả các con cháu của nó sẽ cư xử tốt, là một khái niệm còn xa mới thành hiện thực.

Kurzweil không dành nhiều thời gian cho khái niệm AI thân thiện. "Chúng ta không thể chỉ nói: chúng tôi sẽ đưa cái mã phần mềm nhỏ này vào các AI, và nó sẽ làm cho chúng an toàn," ông nói. "Ý tôi là, điểm mấu chốt ở chỗ các mục tiêu và ý định của trí tuệ nhân tạo sẽ là gì. Chúng ta đang đối mặt với các thử thách làm nản lòng người."[6]

Thật đau đầu khi nghĩ về AI *không* thân thiện – AGI được thiết kế với mục đích tiêu diệt kẻ thù, một hiện thực mà chúng ta sẽ phải sớm đối diện. "Tại sao lại có một thứ như thế?" Kurzweil hỏi. Bởi hàng tá tổ chức ở Mỹ sẽ thiết kế và chế tạo nó, các kẻ thù của Mỹ cũng vậy. Nếu AGI đã tồn tại ngày hôm nay, tôi không hề nghi ngờ chuyện nó sẽ sớm được đưa vào các loại robot chiến binh. DARPA có thể khăng khăng rằng không có gì phải lo lắng — AI được DARPA tài trợ sẽ chỉ giết các kẻ thù của chúng ta. Những người sản xuất nó sẽ cài đặt các khóa an toàn, chế độ an toàn, công tắc sát thương và các dấu hiệu bí mật. Chúng sẽ kiểm soát siêu trí tuệ nhân tạo.

Tháng 12/2011, một người Iran với chiếc laptop chạy một chương trình chia sẻ file đơn giản đã làm hỏng một drone Sentinel. Tháng 7/2008, một cuộc tấn công mạng nhắm vào Lầu Năm góc đã cho các kẻ tấn công quyền truy cập vào 24.000 tài liệu tối mật. Cựu Thứ trưởng Quốc phòng William J. Lynn III kể với tờ *Washington Post* rằng hàng trăm cuộc tấn công mạng vào Bộ Quốc phòng và các nhà thầu của nó đã khiến họ bị mất "những hệ thống nhạy cảm nhất, bao gồm các kỹ thuật điện tử hàng không, các công

nghệ theo dõi, các hệ thống liên lạc vệ tinh và các giao thức bảo mật mạng."[7] Siêu trí tuệ nhân tạo sẽ không bị vô hiệu hóa nhờ những người không làm nổi một việc đơn giản là phòng thủ thành công trước các hacker con người.

Tuy nhiên, chúng ta có thể thu được một số nhận thức quan trọng từ lịch sử kiểm soát vũ khí. Kể từ sự ra đời của vũ khí hạt nhân, chỉ có Mỹ là đã dùng nó để chống lại một kẻ thù. Sức mạnh hạt nhân đã thành công trong việc né tránh tình huống Đảm bảo tiêu diệt lẫn nhau. Không quả bom hạt nhân nào bị phát nổ tình cờ, theo những gì chúng ta đã biết. Hồ sơ của việc quản lý hạt nhân là tốt (tuy mối nguy này vẫn còn tiếp diễn). Nhưng tôi có quan điểm thế này. Quá ít người biết rằng chúng ta cần phải có một cuộc thảo luận quốc tế về AGI, tương tự như những gì đã làm với vũ khí hạt nhân. Quá nhiều người nghĩ rằng biên giới của AI được vạch ra bởi những thứ vô hại như công nghệ tìm kiếm trên mạng, điện thoại thông minh, và bây giờ là Watson. Nhưng AGI thì gần với vũ khí hạt nhân hơn các trò chơi video nhiều.

AI là một công nghệ "kép," một thuật ngữ dược dùng để mô tả các công nghệ ứng dụng trong cả dân sự và quân sự. Ví dụ, phản ứng phân hạch có thể thắp sáng hoặc hủy diệt các thành phố (hoặc cả hai trong trường hợp của Chernobyl và Fukushima Daiichi). Tên lửa được phát triển trong cuộc chạy đua thám hiểm vũ trụ đã tăng cường sức mạnh và độ chính xác của các loại tên lửa đạn đạo vượt đại châu. Công nghệ nano, công nghệ sinh học và kỹ thuật gen đều chứa đựng những triển vọng khổng lồ trong việc ứng dụng dân sinh, nhưng tất cả đều tiềm ẩn các hiểm họa khôn lường và có thể bị lạm dụng trong quân đội và khủng bố.

Khi Kurzweil nói mình là một người lạc quan, ông không ngụ ý rằng AGI sẽ chứng tỏ nó vô hại. Ý ông là sẽ phó mặc cho quá trình điều chỉnh tự nhiên của con người khi họ có các công nghệ nguy hiểm trong tay. Và đôi lúc con người sẽ phải chịu hậu quả.

"Có nhiều lời bàn luận về hiểm họa diệt vong," Kurzweil nói. "Tôi sợ rằng các giai đoạn đau khổ còn dễ xảy ra hơn. Bạn biết đấy, 60 triệu người đã chết trong Thế chiến II. Các công cụ hủy diệt khủng khiếp mà chúng ta có lúc đó hiển nhiên đã làm trầm trọng số thương vong. Tôi tương đối lạc quan rằng chúng ta sẽ vượt qua được chuyện này. Tôi ít lạc quan hơn trong việc chúng ta có thể né tránh các thời kỳ đau khổ."[8]

"Kể từ ngàn xưa khi bắt đầu có lửa, đã có câu chuyện về sự không thể tách rời giữa triển vọng và hiểm họa. Lửa làm chín thức ăn của chúng ta nhưng cũng đốt cháy làng mạc. Bánh xe đã được dùng vào việc tốt, việc xấu và mọi thứ ở giữa. Công nghệ là sức mạnh, và chính công nghệ đó có thể được dùng vào những mục đích khác nhau. Dưới trần gian, con người làm mọi thứ kể từ ân ái cho đến chiến tranh, và chúng ta sẽ tăng cường mọi hoạt động đó với những công nghệ mà chúng ta đã có, và chuyện đó sẽ còn tiếp diễn."

Sự bất định là không thể tránh khỏi, và các tai nạn thì dễ xảy ra – khó mà phản bác điều này. Nhưng sự so sánh là khập khiễng – AI cao cấp hoàn toàn không giống với lửa hoặc bất cứ công nghệ nào khác. Nó sẽ có khả năng suy nghĩ, lập kế hoạch, và biến những nhà chế tạo ra nó thành đồ chơi. Không có bất kỳ một công cụ nào làm được những chuyện như thế. Kurzweil tin rằng: có một cách để giới hạn những khía cạnh nguy hiểm của AI, đặc biệt là ASI, đó là hợp nhất nó với con người qua sự tăng cường trí thông minh – IA. Ngồi trên chiếc ghế bằng kim loại không thoải mái, con

người lạc quan đó nói: "Như tôi đã chỉ ra, AI mạnh sẽ nảy sinh từ nhiều cố gắng khác nhau và sẽ được hợp nhất một cách sâu sắc với cơ sở hạ tầng của nền văn minh con người. Thực vậy, nó sẽ được gắn chặt vào cơ thể và bộ não chúng ta. Vì thế, nó sẽ phản ánh những giá trị của chúng ta bởi nó sẽ là chúng ta."

Và theo lập luận trên thì nó sẽ có độ "an toàn" giống như chúng ta vậy. Nhưng, như tôi đã nói với Kurzweil, *Homo sapiens* không hề đặc biệt vô hại khi giao tiếp với nhau, với các động vật khác hoặc với môi trường sống.

Ai có thể tin rằng con người với bộ não được tăng cường trí thông minh sẽ trở nên thân thiện và nhân từ hơn các máy móc siêu thông minh? Một người được tăng cường, được những người muốn trở thành như vậy gọi là biến nhân, có thể sẽ né tránh các vấn đề Động lực AI cơ bản của Omohundro. Nghĩa là, nó có thể sẽ tự nhận thức và tự cải tiến, nhưng cũng tự xây dựng trong nó một bộ quy tắc đạo đức mang tính nhân bản, thứ sẽ thay thế cho các động lực cơ bản mà Omohundro đã suy ra từ hình mẫu kinh tế tác nhân duy lý. Tuy nhiên, không giống như câu chuyện *Flowers for Algernon* (Hoa trên mộ Algernon)◎, chúng ta không hề biết chuyện gì sẽ xảy ra đối với đạo đức của một người khi trí thông minh của anh ta được tăng lên mức siêu nhân. Có rất nhiều ví dụ về những người có trí thông minh trung bình đã gây chiến với gia đình, trường học, công việc và nơi họ sống. Các thiên tài cũng có khả năng tạo ra bạo động – trong phần lớn các trường hợp, các tướng lĩnh trên thế giới không phải là những kẻ ngu đần. Siêu trí tuệ nhân tạo rất có thể sẽ là một thứ khiến bạo lực leo thang. Nó có thể biến hận thù thành giết chóc, bất đồng quan điểm thành tai họa, như cách mà sự hiện diện của một khẩu súng có thể biến một vụ

đánh lộn trở thành vụ giết người. Chúng ta đơn giản là không thể biết. Tuy vậy, sự tăng cường trí thông minh bằng ASI có một sự hiếu chiến được đặt trên cơ sở nguồn gốc sinh vật mà máy móc không có. Con người là một giống loài với bề dày lịch sử của việc tự bảo vệ, chiếm đoạt tài nguyên, chém giết công khai, và cả các động lực khác, những gì mà chúng ta chỉ mới giả định là có trong các máy móc tự nhận thức.

Và ai sẽ là người đầu tiên "hưởng lợi" từ thứ IA đó? Người giàu nhất? Chúng ta vẫn tin rằng giàu không có nghĩa là xấu, nhưng một nghiên cứu gần đây từ Đại học California tại Berkeley đề xuất điều ngược lại. Các thí nghiệm chỉ ra rằng những người thuộc giới thượng lưu giàu có sẽ "biểu thị các khuynh hướng ra quyết định vô đạo đức, chiếm đoạt tài sản của người khác, nói dối trong các cuộc thương lượng, gian lận để tăng khả năng giành giải thưởng, và tán thành cách cư xử bẩn thỉu ở công sở…"[9] Không thiếu gì các ví dụ về những giám đốc và chính trị gia có tiếng, có quá trình thăng tiến dường như gắn liền với sự tha hóa đạo đức, nếu họ có đạo đức. Liệu những chính trị gia hay các doanh nhân có trở thành những người đầu tiên được tăng cường trí thông minh?

Hay những người đầu tiên sẽ là quân nhân? DARPA đang có vai trò lớn trong câu chuyện này, vậy không có gì ngạc nhiên nếu IA sẽ được áp dụng đầu tiên trên chiến trường hoặc ở Lầu Năm góc. Và DARPA sẽ muốn đòi lại tiền của nó nếu siêu trí tuệ nhân tạo làm cho các binh sĩ cực kỳ thân thiện.

IA *có thể* sẽ xảy đến trong một tương lai khi có những cách tốt hơn bây giờ để kiểm soát nó, với những biện pháp an toàn mà hiện chúng ta chưa tưởng tượng ra được. Có nhiều ASI có vẻ sẽ an toàn hơn là chỉ có một. Còn tốt hơn thế nếu ta có cách nào đó để theo dõi và quan sát AI, nhưng

nghịch lý ở chỗ có lẽ thứ "tập sự" tốt nhất cho điều đó lại là AI. Chúng ta sẽ khảo sát các phương pháp phòng thủ trước ASI ở Chương 14. Tóm lại, IA không phải là giải pháp an toàn về đạo đức. Siêu trí tuệ nhân tạo có thể nguy hiểm chết người hơn bất cứ thứ vũ khí hoặc công nghệ nào hiện đang được kiểm soát nghiêm ngặt nhất.

Chúng ta sẽ phải phát triển song song với IA một khoa học để chọn lựa những ứng viên được tăng cường trí thông minh. Sự tự phụ của những người theo thuyết Singularity khi cho rằng bất cứ ai có đủ tiền trả đầu sẽ được hưởng siêu trí tuệ nhân tạo qua con đường tăng cường bộ não, cũng đồng nghĩa với việc tất cả những người khác sẽ phải sống dưới ách thống trị của siêu trí tuệ nhân tạo độc ác đầu tiên đạt được bằng cách này.[◎] Đó là vì như chúng ta đã thảo luận, trong công cuộc phát triển AGI, ưu thế đi đầu mang tính quyết định. Người nào đạt tới AGI đầu tiên sẽ có khả năng tạo ra những điều kiện cần thiết cho một sự bùng nổ trí thông minh. Họ sẽ làm việc đó vì họ sợ rằng những đối thủ cạnh tranh chính của mình, dù là các công ty hay quân đội, sẽ làm điều tương tự, và họ không biết những đối thủ đó đã đến gần đích tới mức nào. Tôi nhận thấy có một khoảng cách khổng lồ giữa các nhà chế tạo AI và AGI, với nghiên cứu về mối nguy mà họ sẽ gặp phải. Một thiểu số các nhà chế tạo AGI mà tôi đã nói chuyện chưa từng đọc bất cứ công trình nào của MIRI, của Viện Tương lai nhân loại, của Viện Đạo đức và Công nghệ mới, hoặc của Steve Omohundro. Nhiều người không biết là có một cộng đồng đang lớn mạnh gồm những người quan ngại về sự phát triển của trí thông minh siêu nhân, những người đã thực hiện các nghiên cứu quan trọng để tiên liệu những hiểm họa khôn lường của nó. Trừ phi nhận thức này thay đổi, tôi tin chắc rằng họ sẽ để

AGI nhảy vọt lên ASI mà không có các biện pháp đề phòng thảm họa kèm theo.

Sau đây là một ví dụ rõ ràng về chuyện đó. Vào tháng 8/2009 ở California, Hội Vì sự tiến bộ của trí tuệ nhân tạo (Association for the Advancement of Artificial Intelligence: AAAI) mời một nhóm cùng thảo luận về mối lo ngại đang lớn dần lên trong công chúng về chuyện các robot làm phản, sự riêng tư bị xâm phạm và các phong trào công nghệ có hơi hướng tôn giáo.

"Một thứ mới mẻ đã xuất hiện trong khoảng năm đến tám năm gần đây," Eric Horvitz, nhà nghiên cứu nổi tiếng của Microsoft, người chủ trì cuộc họp, phát biểu. "Các nhà công nghệ đang đưa ra những viễn cảnh gần như mang tính tôn giáo, và theo một cách nào đó, các ý tưởng của họ cộng hưởng với ý tưởng tương tự trong sách Khải huyền… Tôi có cảm giác rằng sớm hay muộn chúng ta sẽ phải đưa ra một tuyên bố hoặc đánh giá nào đó, đáp lại mối quan ngại của những người đam mê công nghệ và những người rất quan tâm đến sự ra đời của máy móc thông minh."[11] Bất chấp lời hứa hẹn đó, cuộc gặp này là một cơ hội bị bỏ phí. Nó không rộng mở với công chúng hoặc báo chí, chẳng có nhà đạo đức học máy tính và nhà tư tưởng làm về đánh giá rủi ro nào đến cả. Chỉ có các nhà khoa học máy tính được mời tham gia các thảo luận. Chuyện này gần giống với việc yêu cầu các tay đua xe đặt ra giới hạn tốc độ trong nội thị. Một nhóm nhỏ bàn về Ba định luật về robot của Isaac Asimov, một dấu hiệu cho thấy cuộc thảo luận đạo đức này đã không có thông tin gì về một lượng lớn các công trình nghiên cứu đã vượt xa những khái niệm khoa học viễn tưởng đó. Hội thảo nhỏ hẹp của Horvitz đưa ra báo cáo thể hiện lòng hoài nghi về một sự bùng nổ trí thông minh, về Singularity, và về khả năng mất kiểm soát

các hệ thống thông minh. Dù sao thì hội thảo cũng yêu cầu các nhà đạo đức học và tâm lý học tiếp tục nghiên cứu thêm, và chỉ rõ sự nguy hiểm của các hệ thống máy tính có độ phức tạp và không thể tiên liệu ngày càng tăng, bao gồm "… các hành vi nguy hại và không lường trước được của các hệ thống ra quyết định tự động hoặc bán tự động."[12] Tom Mitchell của Đại học Carnegie Mellon, người được DARPA tài trợ để tạo ra kiến trúc nhận thức chung (và AGI tiềm năng) được gọi là NELL, tuyên bố hội thảo này đã thay đổi quan điểm của ông. "Trước đó tôi rất lạc quan về tương lai của AI và nghĩ rằng Bill Joy và Ray Kurzweil tiên đoán không chính xác. Cuộc họp này làm tôi muốn phát ngôn nhiều hơn về các vấn đề đó."[13]

Trong *The Singularity is Near,* Kurzweil đưa ra vài giải pháp cho vấn đề AI chạy trốn. Các giải pháp yếu ớt một cách đáng ngạc nhiên, đặc biệt là lại đến từ một người phát ngôn, một người vô hình trung đang độc quyền thuyết giáo về siêu trí tuệ nhân tạo. Nhưng ở một góc độ khác thì chúng không có gì đáng ngạc nhiên cả. Như tôi đã nói, có một xung đột không thể hóa giải giữa những người cuồng si sự bất tử với bất cứ thứ gì có khả năng giảm tốc, thử thách hoặc cản trở sự phát triển của những công nghệ sẽ biến giấc mơ của họ thành hiện thực. Trong các cuốn sách và bài giảng của mình, Kurzweil chỉ dùng một phần rất nhỏ sự nhạy bén của ông khi nói về sự nguy hiểm của AI và đưa ra rất ít giải pháp, thế nhưng ông vẫn nói rằng mình đã nghiên cứu kỹ vấn đề này. Tại thành phố New York, trong căn phòng thay đồ chật chội, ngoài cửa là một đội làm phim đang húng hắng ho, tôi tự hỏi, chúng ta nên chờ đợi gì từ chỉ một người? Có nên chờ đợi Kurzweil thấu hiểu những triển vọng *và* hiểm họa của Singularity rồi giảng giải cho chúng ta từng ly từng tí? Hay cá nhân ông phải khảo sát xa hơn những thành ngữ kiểu như "bản chất hai mặt không thể rút gọn hơn của

công nghệ" và thấu hiểu thứ triết học sinh tồn được những người như Yudkowsky, Omohundro hay Bostrom xây dựng?

Không, tôi không nghĩ thế. Đây là một vấn đề mà tất cả chúng ta phải cùng nhau đối mặt, với sự trợ giúp của các chuyên gia.

11

Cất Cánh Nhanh

Dù sao thì ngày qua ngày, máy móc đang tiến gần chúng ta hơn; ngày qua ngày, chúng ta đang trở nên quỵ lụy trước chúng; ngày càng nhiều người bị ràng buộc vào chúng như những nô lệ, ngày càng nhiều người cống hiến sức lực cả cuộc đời mình cho sự phát triển của cuộc đời máy móc. Sự trỗi dậy đơn giản chỉ là vấn đề thời gian, rồi thời khắc đó sẽ tới khi những máy móc chiếm lấy quyền lực thực sự trên thế giới và sự hiện tồn của chúng là thứ mà không một ai có tư duy triết học thực sự lại có thể nghi ngờ, dù chỉ trong khoảnh khắc.

— **Samuel Butler**
nhà văn, nhà thơ Anh thế kỷ 19

Hơn lúc nào hết trong lịch sử, nhân loại đang đối diện với một ngã rẽ. Một đường dẫn tới đau khổ và tuyệt vọng sâu sắc, đường còn lại dẫn tới diệt vong hoàn toàn, cầu cho chúng ta được sáng suốt để chọn lựa chính xác.

— **Woody Allen**
đạo diễn nổi tiếng người Mỹ

Cái cách mà I. J. Good khám phá ra sự bùng nổ trí thông minh không khác gì với cách Ngài Isaac Newton khám phá ra lực hấp dẫn. Tất cả những gì ông đã làm là quan sát một sự kiện mà ông tin là vừa không thể tránh khỏi, vừa thực sự tích cực vì nhân loại chắc chắn sẽ tìm ra một loại "trí tuệ cực thông minh" mà con người cần có để giải quyết các vấn đề quá khó đối với mình. Rồi sau khi sống thêm ba thập niên nữa, Good đã đổi ý. Ông nói, chúng ta sẽ tạo ra các máy móc siêu thông minh theo hình hài của mình, và chúng sẽ hủy diệt chúng ta. Tại sao? Bởi cùng một lý do khiến chúng ta sẽ không bao giờ đồng ý ra lệnh cấm nghiên cứu AI, và cũng lý do đó khiến chúng ta nhiều khả năng sẽ trao tự do cho Đứa trẻ Bận rộn. Cũng chính vì lý do đó mà nhà chế tạo AI Steve Omohundro, một người hoàn toàn thấu suốt, và mọi chuyên gia AI khác mà tôi đã gặp, đều tin rằng việc ngăn chặn phát triển AGI cho đến khi chúng ta biết thêm về các mối nguy của nó là điều không thể thực hiện được.

Chúng ta sẽ không dừng việc phát triển AGI lại, vì nỗi sợ các AI nguy hiểm còn thấp hơn nỗi sợ các quốc gia khác trên thế giới sẽ tiếp tục nghiên cứu AGI, bất chấp cộng đồng quốc tế có nói hoặc làm gì đi nữa. Chúng ta tin rằng khôn ngoan hơn cả là nên ra tay trước. Chúng ta đang ở trong cuộc chạy đua trí thông minh, và không may cho nhiều người, nó đang định hình thành một cuộc cạnh tranh toàn cầu còn ghê gớm hơn cả thứ mà chúng ta vừa thoát khỏi: cuộc chạy đua vũ khí hạt nhân. Chúng ta sẽ cúi đầu để những kẻ hoạch định chính sách và những kẻ tiên phong công nghệ dẫn chúng ta tới địa ngục, theo lời Good, "như loài chuột."

Singularity tích cực của Ray Kurzweil không cần một sự bùng nổ trí thông minh – Định luật Hồi quy Tăng tốc đảm bảo sự tăng trưởng theo hàm mũ liên tục của các công nghệ thông tin, bao gồm những thứ thay đổi

thế giới như AGI, và sau đó là ASI. Xin nhắc lại rằng AGI là một điều kiện cần của sự bùng nổ trí thông minh kiểu Good. Sự bùng nổ này sẽ tạo ra trí thông minh siêu nhân hay ASI. Kurzweil cho rằng sẽ chinh phục được AGI, lúc đầu chậm, và sau đó nhanh chóng hoàn tất, nhờ sức mạnh của LOAR.

Kurzweil không quan tâm tới các vật cản trên đường tới AGI, vì ông tin vào cách dùng kỹ nghệ đảo ngược bộ não. Ông tin rằng chẳng có gì trong bộ não, thậm chí cả ý thức, là không thể đạt được bằng điện toán. Rất ít người tin rằng sự bùng nổ trí thông minh theo kiểu Good là cần thiết để đạt tới ASI sau khi đã chinh phục được AGI. Một sự phát triển chậm chắc là đủ, nhưng như Kurzweil đã nhấn mạnh, có lẽ sẽ không phải chậm và chắc, mà là nhanh và tăng tốc.

Tuy nhiên, sự bùng nổ trí thông minh có thể là không tránh được khi một hệ thống AGI nào đó xuất hiện. Khi một hệ thống nào đó trở nên tự nhận thức và tự cải tiến, thì các động lực cơ bản của nó, như Omohundro đã mô tả, gần như đảm bảo rằng nó sẽ tìm kiếm sự tự cải tiến, hết lần này đến lần khác.◎

Vậy, sự bùng nổ trí thông minh chắc chắn sẽ xảy ra? Hay có thứ gì đó sẽ cản nó lại?

Những người không tin vào AGI xoay quanh hai ý: kinh tế và sự phức tạp của phần mềm. Thứ nhất là kinh tế, họ cho rằng ngân sách sẽ không đủ để đưa AI hẹp trở thành các kiến trúc nhận thức mạnh và có độ phức tạp cao hơn nhiều như AGI. Chỉ một số ít nỗ lực về AGI được tài trợ tốt. Điều này nhắc tới một số nhà nghiên cứu luôn cảm thấy lĩnh vực của họ bị trì hoãn vô thời hạn, một thứ "mùa đông AI." Họ sẽ thoát khỏi đó nếu chính phủ hoặc một công ty lớn như IBM hoặc Google cho rằng AGI là thứ ưu

tiên hàng đầu, và lập ra một dự án cỡ như Manhattan hòng đạt tới nó. Trong Thế chiến II, việc đẩy mạnh nghiên cứu vũ khí hạt nhân đã tiêu tốn của chính phủ Mỹ khoảng hai tỉ đô-la theo thời giá hiện nay, và tuyển dụng khoảng 1.300 người. Những nhà nghiên cứu muốn *sớm* đạt được AGI thường xuyên nhắc tới Dự án Manhattan. Vậy ai sẽ muốn làm chuyện đó, và tại sao?

Những người dùng lý do sự phức tạp của phần mềm thì cho rằng vấn đề AGI đơn giản là quá khó với con người, dù chúng ta có cố gắng đến đâu đi nữa. Như triết gia Daniel Dennett đề xuất, có lẽ chúng ta không sở hữu được thứ trí tuệ có thể hiểu được trí tuệ của chính chúng ta. Trí tuệ của nhân loại có lẽ không phải là thứ mạnh nhất. Nhưng chắc là cần phải có một trí thông minh lớn hơn thế để thấu hiểu hoàn toàn trí khôn của con người.

• • •

Để thăm dò tính hợp lý của những người không tin vào sự bùng nổ trí thông minh, tôi tới gặp một người mà tôi luôn thấy ở các hội thảo về AI, và tôi cũng thường đọc trên web các blog, bài báo, tiểu luận của anh. Anh là một nhà chế tạo AI, đã xuất bản nhiều bài luận và phỏng vấn, thêm vào đó là *chín* cuốn sách bìa cứng, và vô số bài báo khoa học, nên sẽ chẳng có gì đáng ngạc nhiên nếu tôi khám phá ra có một robot tại nhà anh ở ngoại ô Washington, D.C., nó làm việc cả ngày, viết lách hộ Tiến sĩ Benjamin Goertzel để anh Ben Goertzel có thể đi hội thảo. Kết hôn hai lần và có ba đứa con, anh từng làm việc ở những khoa về khoa học máy tính, toán học và tâm lý ở các trường đại học ở Mỹ, Úc, New Zealand và Trung Quốc. Anh là người tổ chức hội thảo quốc tế thường niên duy nhất về trí tuệ nhân

tạo phổ quát, và góp phần khiến thuật ngữ AGI trở nên thông dụng hơn bất kỳ ai khác. Anh là CEO của hai công ty công nghệ, trong đó Novamente là một trong số ít công ty được các chuyên gia kỳ vọng sẽ giải mã được AGI đầu tiên.

Nói một cách tổng quát, kiến trúc nhận thức của Goertzel, có tên OpenCog, biểu thị một cách tiếp cận thiên về kỹ thuật, khoa học máy tính. Những nhà nghiên cứu theo phương pháp khoa học máy tính muốn thiết kế AGI với kiến trúc hoạt động tương tự bộ não người, theo như các môn khoa học nhận thức đã mô tả.[◎] Chúng bao gồm ngôn ngữ học, tâm lý học, nhân loại học, giáo dục, triết học, và nhiều nữa. Các nhà nghiên cứu khoa học máy tính tin rằng chế tạo trí thông minh y hệt như cách bộ não làm việc – kỹ nghệ đảo ngược bản thân bộ não như cách Kurzweil và những người khác tiến cử — là phí phạm thời gian không cần thiết. Hơn thế, thiết kế của bộ não là không tối ưu – lập trình sẽ tạo ra thứ tốt hơn. Họ lý luận, nói cho cùng thì con người không cần phải dùng kỹ nghệ đảo ngược con chim để học cách bay. Từ việc quan sát các con chim và thử nghiệm, họ suy ra các nguyên lý về việc bay. Các khoa học nhận thức là "nguyên tắc bay" của bộ não.

Chủ đề cấu thành OpenCog là trí thông minh có cơ sở từ sự nhận thức các hình mẫu ở mức độ cao. Thường thì các "hình mẫu" trong AI là những khối dữ liệu (tệp tin, hình ảnh, đoạn ký tự, mục đích) đã được phân loại – được tổ chức thành các thư mục – hoặc sẽ bị một hệ thống đã được huấn luyện về dữ liệu phân loại. Các bộ lọc "thư rác" trong email của bạn là một thứ phát hiện hình mẫu trình độ cao – nó phát hiện một hoặc nhiều dấu hiệu của một email không mong muốn (ví dụ, cụm từ "tráng dương" trong tiêu đề email) và lọc nó ra.

Ý niệm về nhận thức hình mẫu của OpenCog thì tinh tế hơn. Hình mẫu mà nó tìm được trong mỗi thứ hoặc mỗi ý tưởng là một chương trình nhỏ, chứa đựng một dạng mô tả của thứ đó. Đây là cách máy tính hiểu về *khái niệm*. Ví dụ, khi bạn nhìn thấy một con chó, ngay lập tức bạn hiểu nhiều điều về nó – bạn vốn đã có một *khái niệm* về chó trong trí nhớ của mình. Mũi nó ướt, nó thích thịt lợn xông khói, nó rụng lông và nó đuổi mèo. Rất nhiều thứ được gói gọn trong khái niệm của bạn về một con chó.

Khi bộ cảm biến của OpenCog thấy một con chó, *chương trình* "chó" của nó sẽ chạy ngay tức thì, tập trung sự chú ý của OpenCog về khái niệm chó. OpenCog sẽ đưa thêm các thông tin vào khái niệm chó dựa trên những chi tiết của con chó đó, hoặc bất kỳ con chó nào khác.

Những module riêng biệt của OpenCog sẽ thực hiện các nhiệm vụ như nhận dạng, tập trung chú ý và ghi nhớ. Chúng làm những việc đó bằng bộ công cụ phần mềm về lập trình di truyền và mạng neuron, những thứ quen thuộc nhưng đã được tùy chỉnh.

Rồi quá trình học bắt đầu. Goertzel dự định sẽ "nuôi lớn" AI trong một thế giới ảo do máy tính giả lập, ví dụ như Second Life, một quá trình học dần thấm lâu có thể sẽ kéo dài nhiều năm. Cũng như những người đang xây dựng các kiến trúc nhận thức khác, Goertzel tin rằng trí thông minh cần phải được "cơ thể hóa," theo một kiểu "tương đối giống con người," kể cả khi cơ thể của nó chỉ tồn tại trong một thế giới ảo. Sau đó tác nhân thông minh còn non dại này sẽ có thể phát triển một tập hợp các sự thật về thế giới mà nó đang sống. Trong giai đoạn học hỏi của nó, thứ được Goertzel xây dựng phỏng theo lý thuyết về sự phát triển của trẻ em của nhà tâm lý học Jean Piaget, em bé OpenCog có thể tăng cường hiểu biết bằng cách

truy cập vào một trong nhiều kho dữ liệu kiến thức phổ thông mang tính thương mại.

Một kho kiến thức khổng lồ như vậy có tên là Cyc, viết tắt của *encyclopedia*. Được công ty Cycorp tạo dựng, nó chứa khoảng một triệu thuật ngữ, năm triệu quy tắc và mối quan hệ thực tế giữa những thuật ngữ đó. Một người sẽ phải mất hơn 1.000 năm để nhập lượng kiến thức này vào theo dạng logic bậc nhất, một hệ thống chính thức được sử dụng trong toán học và khoa học máy tính để biểu thị các khẳng định và mối quan hệ. Cyc không khác gì một cái giếng khổng lồ chứa những kiến thức sâu sắc của nhân loại – nó "hiểu" khá nhiều về tiếng Anh, đến khoảng 40%. Ví dụ, Cyc "biết" cái cây nghĩa là gì, và nó biết cái cây thì có rễ. Nó cũng biết các gia đình thì có tổ tiên *và* gia phả©. "Nó biết rằng việc đặt báo dài hạn sẽ dừng lại khi người ta chết, và cốc chén thì có thể dựng hoặc rót chất lỏng ra, nhanh hoặc chậm.

Trên tất cả, Cyc có một cơ cấu "suy luận." Suy luận là khả năng đưa ra kết luận từ bằng chứng. Cơ cấu suy luận của Cyc hiểu các câu hỏi và đưa ra được các câu trả lời từ kho dữ liệu kiến thức rộng lớn của nó.

Được nhà tiên phong về AI là Douglas Lenat xây dựng, Cyc là dự án AI lớn nhất trong lịch sử, và có lẽ là dự án được tài trợ tốt nhất, với 50 triệu đô-la vốn đến từ các cơ quan nhà nước, trong đó có DARPA, kể từ năm 1984.[3] Những người xây dựng Cyc tiếp tục cải tiến cơ sở dữ liệu và cơ cấu suy luận của nó để nó có thể xử lý tốt hơn thứ "ngôn ngữ tự nhiên," tức ngôn ngữ đọc viết hàng ngày. Một khi nó đã đạt đến khả năng xử lý ngôn ngữ tự nhiên (natural language processing: NLP) đủ tốt, các nhà chế tạo ra nó sẽ bắt đầu cho nó đọc và hiểu tất cả các trang web trên Internet.

Một đối thủ cạnh tranh khác về ngôi vị cơ sở dữ liệu kiến thức thông thái nhất đã bắt đầu làm việc đó. Hệ thống NELL (Never Ending Language Learning: Học ngôn ngữ không ngừng) của Đại học Carnegie Mellon, biết hơn 390.000 sự kiện của thế giới này. Chạy 24/7, NELL – một dự án được DARPA tài trợ – quét hàng trăm triệu trang web tìm các hình mẫu dạng văn bản để nó có thể học nhiều hơn nữa. Nó phân loại các dữ kiện vào 274 thư mục như thành phố, người nổi tiếng, thực vật, các đội thể thao… Nó biết về những dữ kiện liên thư mục, ví dụ như Miami là một thành phố, có đội Miami Dolphins chơi bóng bầu dục. NELL có thể *suy luận* được rằng *dolphins* ở đây không phải là loài động vật có vú sống thành đàn nơi đại dương dù cùng tên.

NELL tận dụng ưu thế có được từ mạng lưới trí lực không chính thống của Internet – mạng lưới người dùng. CMU mời công chúng lên mạng và giúp đào tạo NELL bằng cách phân tích cơ sở dữ liệu và sửa các lỗi sai của nó.[4]

Kiến thức sẽ là chìa khóa đến AGI, cũng như kinh nghiệm và sự hiểu biết – trí thông minh cấp độ con người không thể đạt được nếu thiếu chúng. Nghĩa là mọi hệ thống AGI sẽ phải gắn liền với việc thu thập kiến thức — dù bằng cách gắn cho nó một cơ thể có khả năng thu lượm kiến thức, hoặc bằng cách truy cập vào các cơ sở dữ liệu, hoặc bằng cách đọc toàn bộ nội dung của web. Và càng sớm càng tốt, Goertzel nói.

Để thúc đẩy tiến độ dự án, lãng tử Goertzel phân chia thời gian của anh giữa Hong Kong và Rockville, Maryland. Vào một buổi sáng mùa xuân, tôi thấy trong sân nhà anh một tấm bạt lò xo đã bạc màu và một chiếc minivan Honda trông tàn tạ đến nỗi như vừa bay qua một vành đai thiên thạch để đến đây. Đằng sau xe là miếng dán ghi “Con tôi là người tù của tháng tại

trại giam hạt." Cùng sống với Goertzel và con gái anh là lũ thỏ, một con vẹt và hai con chó. Những con chó chỉ tuân theo các mệnh lệnh được đưa ra bằng tiếng Bồ Đào Nha – Goertzel sinh ra ở Brazil năm 1966 – để những người khác không ra lệnh được cho chúng.

Vị giáo sư mở cửa đón tôi, anh vừa ra khỏi giường vào 11 giờ sáng sau khi đã thức cả đêm để lập trình. Tôi cho rằng chúng ta không nên định sẵn trong đầu là các nhà khoa học đã đi vòng quanh thế giới phải trông thế nào, vì trong nhiều trường hợp nó chẳng đúng chút nào, chí ít là với tôi. Trên giấy tờ thì Tiến sĩ Benjamin Goertzel khiến ta nghĩ đến một học giả ngành máy tính, cao, gầy, có lẽ hói, theo chủ nghĩa quốc tế đơn thuần, và chắc là vừa nằm vừa đi xe đạp. Trời ạ, chỉ có gầy và chủ nghĩa quốc tế là đúng. Goertzel đời thực trông như một gã hippy quá đà. Nhưng đằng sau cặp kính kiểu John Lennon, bộ tóc dài và gần như bện xoắn lại, bộ râu lúc nào cũng lởm chởm, nụ cười nửa miệng cố định, anh đã diễn giải không mệt mỏi cho tôi về các lý thuyết phức tạp đến chóng mặt, sau đó thì quay lại giải thích nó về mặt toán học. Anh viết quá tốt nên không thể là một nhà toán học thông thường, và làm toán quá giỏi nên không thể là một người viết thông thường. Thế nhưng anh lại quá khôn ngoan, đến nỗi khi nói với tôi rằng anh đã nghiên cứu đạo Phật và chưa hiểu được nhiều, tôi tự hỏi con người an nhiên, thanh thản này trông sẽ thế nào khi anh *hiểu hết*.

Tôi đến để hỏi anh về những tính chất cơ bản của sự bùng nổ trí thông minh và *những tác nhân cản trở* – những vật cản có thể khiến nó không xảy ra. Sự bùng nổ trí thông minh có hợp lý không, và có đúng là không thể tránh khỏi không? Nhưng trước hết, sau khi chúng tôi đã yên vị trong phòng khách với lũ thỏ, anh liền giải thích tại sao anh nghĩ khác với hầu hết những nhà chế tạo và nhà lý thuyết AI khác.

Nhiều người, đặc biệt là những người ở MIRI, khuyên rằng nên dùng thật nhiều thời gian khi phát triển AGI, để có thể hoàn toàn tuyệt đối chắc chắn rằng "tính thân thiện" đã là một phần của nó. Chuyện AGI bị trì hoãn và ra đời chậm hàng thế kỷ làm họ hạnh phúc, vì họ tin tưởng mạnh mẽ rằng siêu trí tuệ nhân tạo sẽ hủy diệt chúng ta. Và có lẽ không chỉ chúng ta, mà tất cả các dạng sống trong Ngân hà.◎

Goertzel thì không. Anh khuyến cáo rằng nên chế tạo AGI càng sớm càng tốt. Năm 2006, anh đã có một bài phát biểu với nhan đề "10 năm để đến được một Singularity tích cực – Nếu chúng ta thực sự, thực sự cố gắng." "Singularity" trong câu này mang định nghĩa được dùng nhiều nhất hiện nay – là thời gian mà con người đạt tới ASI, và chia sẻ Trái đất với một thực thể thông minh hơn chính mình. Goertzel lập luận rằng nếu AGI ra sức lợi dụng ưu thế từ cơ sở hạ tầng của xã hội hoặc nền công nghiệp mà ở đó nó đã sinh ra, và trí thông minh của nó "bùng nổ" thành ASI, thì phải chăng chúng ta nên để vụ "cất cánh nhanh" đó (một sự bùng nổ trí thông minh đột ngột và không kiểm soát được) xảy ra trong thế giới còn mông muội này, thay vì trong một thế giới tương lai khi mà công nghệ nano, công nghệ sinh học và tự động hóa diện rộng sẽ làm AI lên ngôi cực nhanh?

Để tìm được câu trả lời, xin trở lại với Đứa trẻ Bận rộn một lát. Như bạn đã biết, nó đã "cất cánh nhanh" từ AGI lên ASI rồi. Nó đã trở nên tự nhận thức và tự cải tiến, trí thông minh của nó đã vọt lên quá cấp độ con người trong vài ngày. Giờ thì nó muốn ra khỏi siêu máy tính, nơi nó được tạo ra để thỏa mãn các động lực của mình. Như Omohundro đã thảo luận, các động lực gồm: hiệu suất, tự bảo tồn, chiếm đoạt tài nguyên và sáng tạo.[6]

Như chúng ta đã thấy, một ASI không bị quản thúc sẽ biểu hiện các động lực này ra theo những cách hết sức điên cuồng. Để có được những gì nó muốn, nó có thể trở nên thuyết phục một cách khủng khiếp, thậm chí là đáng sợ. Nó sẽ sử dụng sức mạnh trí tuệ vượt trội để tháo dỡ bức tường phòng vệ của Người giữ cửa. Sau đó, bằng cách phát minh và thao túng các loại công nghệ, trong đó có công nghệ nano, nó sẽ đoạt lấy quyền kiểm soát tài nguyên, kể cả các phân tử của chúng ta.

Do đó, Goertzel nói, hãy xem xét cẩn thận cái thế giới mà bạn đưa trí thông minh siêu nhân vào, nó đang có những loại công nghệ gì. Ví dụ, *ngay bây giờ* sẽ an toàn hơn là 50 năm sau.

"Sau 50 năm," anh nói với tôi, "nền kinh tế sẽ trở nên hoàn toàn tự động hóa, với một cơ sở hạ tầng phát triển hơn nhiều. Nếu một máy tính muốn cải tiến phần cứng của nó, thì không cần phải nhờ đến con người. Nó chỉ cần lên mạng và sau đó một số robot sẽ đến và giúp nó cài đặt phần cứng mới. Sau đó nó sẽ ngày càng thông minh hơn, tiếp tục đặt hàng các bộ phận mới và tự lắp ráp bản thân nó, sẽ không ai biết chuyện gì đang xảy ra. Vậy là có lẽ khoảng 50 năm nữa bạn sẽ gặp một siêu AGI, thứ *thực sự có thể* trực tiếp chiếm lấy thế giới. Con đường đi đến quyền lực của AGI sẽ đầy kịch tính."

Vào thời điểm này, hai con chó của Goertzel nhập bọn với chúng tôi, chúng nhận được vài mệnh lệnh bằng tiếng Bồ Đào Nha. Rồi chúng đi ra ngoài sân sau để chơi.

"Nếu anh tin rằng cất cánh nhanh là một thứ nguy hiểm, thì cách an toàn nhất là phát triển AGI cao cấp sớm nhất có thể, để điều đó xảy ra khi các công nghệ hỗ trợ còn yếu và một vụ cất cánh nhanh không kiểm soát được

sẽ khó xảy ra. Và nên cố gắng tạo ra nó trước khi chúng ta có công nghệ nano mạnh hoặc các robot có khả năng tái định hình, tức robot có thể tự thay đổi hình dạng và chức năng để phù hợp với mọi công việc."

Theo một nghĩa rộng, Goertzel không thực sự tin vào ý tưởng một vụ cất cánh nhanh sẽ dẫn đến ngày tận thế – kịch bản Đứa trẻ Bận rộn. Cách anh lập luận rất đơn giản: chúng ta sẽ chỉ tìm ra được cách chế tạo các hệ thống AI có đạo đức qua việc xây dựng chúng, chứ không thể chỉ đứng từ xa mà suy luận rằng chúng sẽ nguy hiểm. Nhưng anh không bỏ qua các hiểm họa.

"Tôi sẽ không nói rằng tôi không lo lắng về nó. Tôi sẽ nói rằng có một sự bất định khổng lồ và không thể giảm bớt trong tương lai. Con gái và con trai tôi, mẹ tôi, tôi không muốn tất cả hộ phải chết vì một số AI siêu nhân tái xử lý các phân tử của họ thành thứ vật chất điện toán. Nhưng tôi nghĩ rằng lý thuyết chế tạo AGI có đạo đức sẽ đến qua việc thử nghiệm các hệ thống AGI."

Khi Goertzel nói thế, quan điểm tuần tự nghe rất có lý. Có một sự bất định khổng lồ và không thể giảm bớt trong tương lai. Và các nhà khoa học sẽ thu được nhiều nhận thức về cách quản lý những máy móc thông minh trên con đường tới AGI. Sau cùng thì chính con người tạo ra máy móc. Máy tính sẽ không trở nên kỳ dị một cách đột ngột khi chúng hóa thành thông minh. Và vì thế, lập luận này tiếp tục, chúng sẽ làm như được bảo. Thực tế, chúng ta thậm chí còn có thể chờ đợi rằng chúng sẽ *đạo đức* hơn cả chúng ta, vì chúng ta đâu có muốn chế tạo một trí thông minh với nỗi thèm khát bạo lực và giết chóc, đúng không nào?

Nhưng đó chính xác là kiểu của các drone tự động và robot trận mạc mà chính phủ Mỹ và các nhà thầu quân đội hiện đang phát triển. Họ đang xây dựng và sử dụng các AI cao cấp nhất sẵn có. Tôi cảm thấy lạ khi người đi đầu trong lĩnh vực robot, Rodney Brooks, lại không cho rằng siêu trí tuệ nhân tạo sẽ có hại khi mà iRobot, công ty do ông sáng lập, đã sản xuất các robot sát thương rồi.[7] Tương tự, Kurzweil lập luận rằng AI cao cấp sẽ mang những giá trị của chúng ta vì nó đến từ chúng ta, và vì thế sẽ vô hại.

Tôi đã phỏng vấn cả hai nhà khoa học từ 10 năm trước, và họ đều lập luận giống nhau. Trong thập niên tiếp theo, họ vẫn giữ sự nhất quán đáng buồn đó, dù tôi nhớ từng nghe một bài nói của Brooks, trong đó ông tuyên bố việc chế tạo robot sát thương có sự khác biệt về đạo đức với quyết định chính trị để sử dụng chúng.

Tôi nghĩ có một khả năng cao rằng sẽ có những sai lầm đau thương trên con đường tới AGI, cũng như khi các nhà khoa học đã thực sự đạt tới nó. Như tôi đã đề xuất, chúng ta sẽ phải gánh chịu hậu quả lâu dài từ rất lâu trước khi có cơ hội học hỏi về chúng, như dự báo của Goertzel. Về khả năng sống sót của chúng ta – tôi tin rằng mình đã thể hiện khá rõ sự nghi ngờ điều đó. Nhưng có thể bạn sẽ ngạc nhiên khi biết vấn đề chủ chốt của tôi với việc nghiên cứu AI thậm chí không phải là điều đó. Vấn đề là có quá nhiều người hiểu rằng sẽ không có *bất kỳ rủi ro nào* trong giai đoạn phát triển AI. Những người sẽ sớm phải chịu đựng những hệ quả xấu từ AI có quyền được biết một số tương đối ít các nhà khoa học đang đưa cả nhân loại đến đâu.

Như tôi đã nói, sự bùng nổ trí thông minh của Good và thái độ bi quan của ông về tương lai nhân loại là một điều quan trọng, vì nếu sự bùng nổ trí thông minh là khả thi, thì xác suất của AI ngoài-tầm-kiểm-soát cũng vậy.

Trước khi xem xét các tác nhân cản trở của nó – kinh tế và sự phức tạp của phần mềm – hãy cùng nhìn lại con đường tới ASI Những thành phần cơ bản của sự bùng nổ trí thông minh là gì?

Đầu tiên, một sự bùng nổ trí thông minh cần có AGI hoặc thứ gì đó rất gần với nó. Tiếp theo, Goertzel, Omohundro và những người khác, đều nhất trí rằng nó sẽ phải có khả năng tự nhận thức – nghĩa là nó phải có một sự hiểu biết sâu sắc về bản thể của nó.[8] Vì nó là một AGI nên chúng ta giả định rằng nó có trí thông minh phổ quát. Nhưng để tự cải tiến, nó cần phải có nhiều hơn thế. Nó sẽ cần những kiến thức lập trình nhất định để khởi phát vòng lặp tự cải tiến, trái tim của sự bùng nổ trí thông minh.

Theo Omohundro, cách thức tự cải tiến và lập trình đến từ sự duy lý của AI – tự cải tiến để theo đuổi các mục tiêu là một cách hành xử duy lý. Không có khả năng tự cải thiện chương trình sẽ là một nhược điểm lớn.[9] AI sẽ bị thúc đẩy phải có được các kỹ năng lập trình. Nhưng làm thế nào nó có thể có chúng? Hãy cùng xem một kịch bản giả định đơn giản với OpenCog của Goertzel.

Kế hoạch của Goertzel là tạo ra một "tác nhân" AI giống như một em bé và thả nó tự do trong một thế giới ảo có kết cấu phong phú để nó học hỏi. Anh sẽ hỗ trợ quá trình học của nó với một cơ sở dữ liệu kiến thức, hoặc cho tác nhân này kỹ năng NLP và cho nó đọc Internet. Các giải thuật học hỏi mạnh, thứ được tạo ra trong tương lai, sẽ biểu thị kiến thức với "các giá trị sự thật xác suất." Nghĩa là sự hiểu biết của tác nhân này về một thứ gì đó có thể được cải thiện nếu có thêm các ví dụ hoặc dữ liệu. Một cơ cấu suy luận xác suất cũng đang được xây dựng sẽ cho nó khả năng lập luận trên các bằng chứng không hoàn chỉnh.[10]

Với lập trình di truyền, Goertzel có thể dạy tác nhân AI này cách tự tiến hóa các công cụ máy học của nó – những chương trình của bản thân nó. Những chương trình này sẽ cho phép tác nhân thử nghiệm và học hỏi – hỏi các câu hỏi đúng về môi trường xung quanh, phát triển các giả thiết và kiểm tra chúng. Những gì nó có thể học sẽ gần như vô hạn. Nếu có thể tiến hóa thành những chương trình tốt hơn, nó có thể cải tiến các giải thuật của nó.

Vậy thì điều gì sẽ ngăn chặn sự bùng nổ trí thông minh xảy ra trong thế giới ảo này? Có lẽ là không gì cả. Chuyện này khiến một số nhà lý thuyết đề xuất rằng Singularity cũng có thể xảy ra trong một thế giới ảo.[11] Nó có khiến những sự kiện này trở nên an toàn hơn chút nào không là một câu hỏi cần được đào sâu. Một cách khác là cài đặt tác nhân thông minh này vào một robot, tiếp tục giáo dục nó và thỏa mãn các mục tiêu của nó trong thế giới thực. Một cách khác nữa là dùng tác nhân AI này để tăng cường trí thông minh cho não người.

Nói chung, những người tin rằng trí thông minh cần được cơ thể hóa thường cho rằng bản thân kiến thức được khởi phát nhờ những kinh nghiệm đến từ cảm giác và vận động. Quá trình xử lý nhận thức không thể xảy ra nếu thiếu chúng. Họ cho là việc học thuộc lòng các dữ kiện về quả táo sẽ không bao giờ khiến bạn thông tuệ theo nghĩa con người về quả táo đó. Bạn sẽ không bao giờ phát triển "khái niệm" về một quả táo bằng cách đọc hoặc nghe về nó – sự hình thành khái niệm đòi hỏi bạn phải ngửi, sờ, nhìn và nếm – càng nhiều càng tốt. Trong AI, điều này được biết đến như "vấn đề nền móng."

Hãy xem xét một số hệ thống với khả năng nhận thức mạnh nằm ở trên AI hẹp nhưng chưa tới được mức AGI. Gần đây, Hod Lipson tại Phòng

nghiên cứu điện toán tổng hợp của Đại học Cornwell đã phát triển một phần mềm có khả năng rút ra được các định luật khoa học từ dữ liệu thô.[12] Bằng cách quan sát sự dao động của một con lắc đôi, nó tái khám phá ra nhiều định luật vật lý của Newton. "Nhà khoa học" ở đây là một giải thuật di truyền. Nó bắt đầu với những giả định thô về các phương trình vận hành con lắc, kết hợp các phần tốt nhất của những phương trình đó, và sau nhiều thế hệ thì cho ra các định luật vật lý, ví dụ như định luật bảo toàn năng lượng.[13]

Và hãy xem xét di sản đáng lo ngại của AM và Eurisko. Đó là những thử nghiệm trước đây của Douglas Lenat, nhà sáng lập Cyc. Sử dụng giải thuật di truyền, AM của Lenat, Nhà toán học tự động (Automatic Mathematician), sinh ra những lý thuyết toán học, về bản chất là tái khám phá các định luật toán học cơ bản bằng cách tạo ra các quy luật từ dữ liệu toán học. Nhưng AM bị giới hạn bởi toán học – Lenat muốn một chương trình giải quyết các vấn đề trong nhiều lĩnh vực chứ không chỉ một. Vào những năm 1980, ông tạo ra Eurisko (tiếng Hy Lạp nghĩa là "Tôi khám phá"). Eurisko đã tạo nên đột phá trong ngành AI bằng việc tiến hóa ra các phương pháp gần đúng, hay các quy luật thực nghiệm, về vấn đề nó đang phải giải quyết, và nó tiến hóa ra các quy luật về bản thân cách hoạt động của nó.[14] Nó rút ra các bài học từ việc giải quyết vấn đề thành công hoặc thất bại, và mã hóa chúng thành những quy luật mới. Nó thậm chí tự sửa đổi chương trình của mình, được viết bằng ngôn ngữ lập trình Lisp.

Thành công lớn nhất của Eurisko là khi Lenat cho nó đấu với con người trong một trò chơi chiến tranh ảo tên là Traveller Trillion Credit Squadron. Trong trò chơi này, người chơi điều khiển các con tàu trong một hạm đội giả định, với một ngân quỹ cho trước, đánh nhau với các hạm đội khác.

Các biến số bao gồm số lượng và loại tàu, độ dày của thân tàu, số lượng và loại súng… Eurisko nghĩ ra một hạm đội, thử nghiệm nó chống lại các hạm đội giả định khác, lấy ra những phần tốt nhất từ bên thắng cuộc và kết hợp chúng lại, thêm vào đó các đột biến, và cứ thế, giống như một phiên bản số của quá trình chọn lọc tự nhiên. Sau 10.000 trận đánh, chạy trên 100 máy tính liên kết với nhau, Eurisko đã tiến hóa ra một hạm đội bao gồm nhiều tàu đứng yên với bộ giáp dày và rất ít vũ khí. Ngược lại, hầu hết người chơi đều chọn các con tàu cỡ trung, tốc độ cao với vũ khí mạnh. Tất cả các đối thủ của Eurisko đều chịu chung số phận – cuộc chơi kết thúc khi tất cả tàu của họ bị tiêu diệt, trong khi một nửa số tàu của Eurisko vẫn hoạt động. Eurisko giành giải thưởng năm 1981 một cách dễ dàng. Năm sau, ban tổ chức thay đổi luật chơi và không công bố trước để Eurisko không thể chơi trước hàng ngàn trận giả lập. Nhưng chương trình này đã rút ra được những quy luật thực nghiệm hiệu quả từ kinh nghiệm trước đó, cho nên nó không cần nhiều vòng lặp. Nó lại thắng một cách dễ dàng. Năm 1983, ban tổ chức trò chơi dọa sẽ giải tán sự kiện này nếu Eurisko lại thắng tiếp lần thứ ba. Lenat rút lui.[15]

Một lần trong quá trình vận hành, Eurisko đã tạo ra một quy luật, nhanh chóng đạt tới giá trị cao nhất, tức là thích hợp nhất. Lenat và đội của ông tìm hiểu cặn kẽ để hiểu quy luật này có gì mà hay vậy. Kết quả là cứ khi nào một giải pháp được đề ra cho một vấn đề giành được sự đánh giá cao, thì quy luật này sẽ gắn tên nó vào giải pháp đó, tự nâng "giá trị" của mình lên.[16] Đây là một khái niệm thông minh nhưng không hoàn chỉnh về giá trị. Eurisko thiếu một sự hiểu biết về mặt ngữ cảnh, rằng trò lách luật nhỏ này không làm nó thắng cuộc được. Đó là lúc Lenat nảy ra ý định sẽ tạo nên một cơ sở dữ liệu khổng lồ về những gì Eurisko còn thiếu – hiểu biết về

đời sống thống thường. Cyc, cơ sở dữ liệu về kiến thức phổ thông mà một người sẽ mất cả ngàn năm để nhập liệu, đã được sinh ra.

Lenat chưa bao giờ công bố mã nguồn của Eurisko, điều đó khiến một số blogger về AI đặt giả thiết rằng ông muốn cho nó hồi sinh vào một ngày nào đó, hoặc ông lo lắng sẽ có ai khác làm điều đó. Đặc biệt là Eliezer Yudkowsky, người đã viết nhiều nhất về hiểm họa AI, nghĩ rằng giải thuật của giai đoạn những năm 1980 là thứ gần nhất với hệ thống AI tự cải tiến mà các nhà khoa học đã tạo ra cho đến nay. Ông kêu gọi những lập trình gia không dùng lại nó.[17]

• • •

Giả định đầu tiên của chúng ta là để sự bùng nổ trí thông minh xảy ra được, hệ thống AGI đang nói tới phải có khả năng tự cải tiến theo kiểu Eurisko và tự nhận thức.

Hãy đặt ra thêm một giả định nữa trong khi chúng ta xem xét những trở ngại và rào cản. Khi trí thông minh của một AI tự nhận thức và tự cải tiến tăng lên, động lực hiệu suất của nó sẽ khiến nó thu mã của mình lại cô đọng nhất có thể, và nén nhiều nhất có thể trí thông minh vào phần cứng của nó. Dù vậy, phần cứng mà nó có thể sử dụng là một yếu tố có giới hạn. Ví dụ, nếu môi trường của nó không đủ dung lượng cho AI tự nhân bản phục vụ cho nhu cầu an ninh và tự cải tiến, thì sao? Tạo ra các vòng lặp tự cải tiến là trái tim của sự bùng nổ trí thông minh của Good. Đây là lý do tại sao đối với kịch bản Đứa trẻ Bận rộn mà tôi đề xuất, sự bùng nổ trí thông minh lại diễn ra trong một siêu máy tính có dung lượng bộ nhớ lớn.

Độ giãn nở của môi trường sống của AI là một yếu tố lớn góp phần vào sự tăng trưởng trí thông minh của nó. Nhưng đây là một vấn đề đơn giản. Thứ nhất, như chúng ta đã biết được từ Định luật Hồi quy Tăng tốc của Kurzweil, tốc độ và dung lượng máy tính nhân đôi trong một khoảng thời gian ngắn là mỗi năm. Điều đó có nghĩa là cho dù một hệ thống AGI có yêu cầu phần cứng thế nào vào hôm nay, thì một năm sau yêu cầu đó sẽ được thỏa mãn với trung bình là *một nửa* phần cứng đó, và với *một nửa* giá.

Thứ hai, khả năng truy cập vào điện toán đám mây. Điện toán đám mây cho phép người dùng thuê sức mạnh và dung lượng máy tính trên Internet. Các công ty như Amazon, Google và Rackspace cung cấp cho người dùng các lựa chọn về tốc độ vi xử lý, hệ điều hành và dung lượng bộ nhớ. Sức mạnh máy tính đã trở thành một dịch vụ thay vì một sự đầu tư phần cứng. Bất cứ ai với một thẻ tín dụng và biết vài thao tác đơn giản đều có thể thuê một siêu máy tính ảo. Ví dụ, trên dịch vụ điện toán đám mây EC2 của Amazon, một nhà cung cấp tên là Cycle Computing đã tạo ra một chùm 30.000 vi xử lý với tên gọi Nekomata (tên một loại linh miêu trong văn hóa Nhật Bản). Cứ 8 vi xử lý trong 30.000 cái thì đi kèm với 7 gigabyte RAM (tương đương với RAM của một máy tính cá nhân), tổng là 26,7 terabyte RAM và 2 petabyte dung lượng ổ cứng (tương đương 40 triệu tủ hồ sơ bốn ngăn).[18] Nhiệm vụ của Nekomata? Mô hình hóa phản ứng phân tử của một hợp chất thuốc mới cho một công ty dược. Đây là một công việc khó ngang với mô phỏng các hiện tượng thời tiết.

Để hoàn thành nhiệm vụ của mình, Nekomata chạy bảy giờ với giá 9.000 đô-la. Trong khoảng đời ngắn ngủi, nó là một siêu máy tính, đứng trong danh sách 500 siêu máy tính nhanh nhất. Nếu một máy tính để bàn làm việc này, nó sẽ phải mất 11 năm. Các nhà khoa học của Cycle

Computing cài đặt mạng đám mây EC2 Amazon từ xa, tại các văn phòng của họ, nhưng phần mềm sẽ quản lý công việc. Như người phát ngôn của công ty đã nói, đó là vì "không có cách nào để người bình thường có thể theo dõi được mọi bộ phận lưu động của chùm máy tính lớn cỡ này."[19] Vậy giả định thứ hai của chúng ta là hệ thống AGI sẽ có đủ không gian để tăng trưởng thành siêu trí tuệ nhân tạo. Thế thì những nhân tố giới hạn sự bùng nổ trí thông minh là gì?

Ta hãy xem xét yếu tố kinh tế trước. Liệu vốn tài trợ cho AGI có mất dần rồi mất hẳn không? Nếu không có doanh nghiệp hoặc chính phủ nào thấy việc chế tạo các máy móc có trí thông minh cấp độ con người là giá trị, hoặc cũng kỳ cục như thế nếu họ nghĩ rằng vấn đề này là quá khó không thực hiện được và không muốn đầu tư, thì sao?

Điều này sẽ đặt các nhà khoa học AGI vào một tình thế khó khăn. Họ sẽ buộc phải bán những yếu tố của các kiến trúc lớn của mình cho các công việc tương đối tầm thường như khai thác dữ liệu hoặc giao dịch chứng khoán. Họ sẽ phải tìm việc làm thêm. Cũng đúng thôi, trừ một số trường hợp ngoại lệ, chuyện này ít nhiều đang là tình trạng hiện nay, và ngay cả thế thì việc nghiên cứu AGI vẫn tiến triển vững chắc.

Hãy xem bằng cách nào mà dự án OpenCog của Goertzel vẫn sống. Những bộ phận thuộc kiến trúc của nó đang hoạt động, bận rộn phân tích các dữ liệu sinh học và giải quyết các vấn đề của mạng lưới điện để kiếm tiền. Lợi nhuận thu được dùng để nghiên cứu và phát triển OpenCog.

Công ty Numenta, đứa con trí tuệ của Jeff Hawkins, tác giả của Palm Pilot và Treo, kiếm sống bằng cách làm việc trong các hệ thống cung cấp điện để lường trước các sự cố.

Trong khoảng một thập niên, Peter Voss phát triển công ty AGI của ông. Adaptive AI, theo kiểu "ẩn danh," ông thuyết giảng rộng rãi về AGI nhưng không hé lộ kế hoạch xử lý nó. Sau đó vào năm 2007, ông thành lập Smart Action, một công ty sử dụng công nghệ của Adaptive AI để xây dựng các Đại lý Ảo. Chúng là các chatbot dùng trong điện thoại dịch vụ khách hàng, có kỹ năng NLP để trợ giúp khách hàng trong các giao dịch mua hàng.

LIDA (Learning Intelligent Distributed Agent: Tác nhân học hỏi điều phối thông minh) của Đại học Memphis có lẽ không cần phải lo lắng nhiều về nguồn vốn để tiếp tục nâng cấp. Là một dạng kiến trúc nhận thức AGI tương tự như OpenCog, một phần vốn tài trợ của LIDA đến từ Hải quân Mỹ. LIDA được dựng trên một kiến trúc (có tên IDA) sử dụng trong hải quân để tìm việc cho những lính thủy sắp giải ngũ. Và khi làm việc đó, "cô ấy" đã thể hiện những kỹ năng nhận thức sơ khai của con người, như bộ phận báo chí của trường đã mô tả dưới đây:

> Cô ấy chọn một số công việc cho một lính thủy, dựa vào những yếu tố như các chính sách của Hải quân, những yêu cầu của công việc, những ưu tiên của lính thủy đó, và những cân nhắc của bản thân cô ấy về các thời hạn khả thi. Sau đó cô ấy thảo luận với lính thủy đó bằng tiếng Anh qua nhiều email về việc chọn lựa công việc. IDA chạy qua một vòng lặp nhận thức trong đó cô ấy cảm nhân các môi trường, nội và ngoại; tạo ra ý nghĩa bằng cách diễn giải môi trường và quyết định xem điều gì là quan trọng; và trả lời câu hỏi duy nhất ở đây [đối với các lính thủy]: "Sắp tới tôi sẽ làm gì?"[20]

Cuối cùng, như chúng ta đã thảo luận ở Chương 3, có nhiều dự án AGI hiện đang ngấm ngầm diễn ra một cách có chủ đích. Những công ty được

gọi là "ẩn danh" đó thường công khai về mục đích, như Adaptive AI của Voss chẳng hạn, nhưng giữ bí mật về công nghệ. Đó là vì họ không muốn tiết lộ công nghệ của mình cho các đối thủ cạnh tranh, những kẻ nhái hàng hoặc trở thành mục tiêu của tình báo. Các công ty ẩn danh khác thì hoạt động ngầm, nhưng không ngại trong việc thu hút các khoản đầu tư. Siri, công ty đã tạo ra trợ lý cá nhân sử dụng công nghệ NLP được đón nhận khá tốt trên iPhone của Apple, được thành lập với cái tên Stealth Company, đúng theo nghĩa đen là công ty ẩn danh. Đây là thông cáo trước khi thành lập trên trang web của họ:

> Chúng tôi sắp thành lập một công ty sẽ trở nên hùng mạnh ở Thung lũng Silicon. Chúng tôi hướng tới việc tái thiết kế một cách cơ bản bộ mặt của thế giới tiêu dùng trên Internet. Chính sách của chúng tôi là ẩn danh, vì chúng tôi sắp bí mật hoàn thiện thứ Vĩ đại Sắp tới. Chúng tôi sẽ tiết lộ câu chuyện của mình một cách kỳ vĩ, sớm hơn những gì bạn nghĩ… [21]

Giờ hãy xét đến vấn đề vốn tài trợ và DARPA, và một câu chuyện kỳ lạ lại dẫn đến Siri.

Từ những năm 1960 cho đến những năm 1990, DARPA đã rót vốn cho nghiên cứu AI nhiều hơn các công ty tư nhân và bất kỳ cơ quan chính phủ nào.[22] Nếu không có tài trợ của DARPA, cuộc cách mạng máy tính có lẽ đã không xảy ra; nếu trí tuệ nhân tạo có phát triển thì cũng phải nhiều năm sau. Trong "thời kỳ vàng son" của AI vào những năm 1960, cơ quan này đầu tư vào nghiên cứu AI cơ bản tại Đại học CMU, MIT, Stanford, và Viện nghiên cứu Stanford. Nghiên cứu AI tiếp tục phát triển hưng thịnh tại

những cơ sở này, và quan trọng là tất cả trừ Stanford ra đều công khai thừa nhận kế hoạch chế tạo AGI, hoặc một thứ rất gần với nó.

Nhiều người biết rằng DARPA (lúc đầu là ARPA) đã tài trợ cho một nghiên cứu phát minh ra Internet (lúc đầu là ARPANET), cũng như đã tài trợ cho những nhà nghiên cứu phát triển một thứ phổ biến hiện nay, GUI, hay Graphical User Interface (Giao diện đồ họa người dùng), phiên bản bạn vẫn nhìn thấy mỗi khi dùng máy tính hay điện thoại thông minh. Thế nhưng cơ quan này cũng là người hậu thuẫn chính của phần mềm và phần cứng xử lý song song, điện toán phân tán, thị giác máy tính, và xử lý ngôn ngữ tự nhiên. Các đóng góp này vào những nền tảng cơ bản của ngành khoa học máy tính cũng quan trọng đối với ngành AI như việc tài trợ hướng đích đặc thù hiện nay của DARPA.

DARPA đang dùng tiền của mình thế nào? Ngân sách hằng năm gần đây phân phối 61,3 triệu đô-la cho hạng mục Máy học, và 49,3 triệu đô-la cho Điện toán nhận thức. Nhưng các dự án AI cũng được tài trợ trong hạng mục Công nghệ thông tin và truyền thông với 400,3 triệu đô-la, và Các chương trình Mật với 107,2 triệu đô-la.[23]

Như được mô tả trong ngân sách của DARPA, các mục tiêu của Điện toán nhận thức là rất tham vọng, y như những gì bạn có thể tưởng tượng.

> Chương trình Các hệ thống điện toán nhận thức… đang phát triển cuộc cách mạng tiếp theo trong điện toán và công nghệ xử lý thông tin, thứ sẽ cho phép các hệ thống điện toán có khả năng suy luận và học hỏi, có trình độ tự động vượt xa những hệ thống hiện nay.
>
> Khả năng tư duy học hỏi và thích nghi sẽ đưa điện toán đến những tầm cao mới về năng lực và cho ra những ứng dụng mới đầy mạnh mẽ. Dự

> án điện toán nhận thức sẽ phát triển các công nghệ nòng cốt giúp các hệ thống điện toán có thể học, lý luận và áp dụng kiến thức thu được qua kinh nghiệm, phản ứng một cách thông minh với những thứ mà nó chưa gặp bao giờ.
>
> Những công nghệ này sẽ khiến các hệ thống biểu đạt nhiều hơn sự tự chủ, khả năng tái cấu hình để tự thích nghi, thương thuyết thông minh, hành xử hợp tác và khả năng sống sót với ít sự can thiệp của con người.[24]

Nếu điều đó nghe có vẻ giống AGI, thì đó là vì có những lý do tốt để tin rằng chính là như vậy. DARPA không tự nghiên cứu và phát triển AI, nó tài trợ để những người khác làm việc đó, cho nên tiền từ ngân quỹ của nó hầu như toàn được rót vào các trường đại học dưới dạng hỗ trợ nghiên cứu. Vì thế, ngoài những dự án AGI mà chúng ta đã nhắc tới, trong đó các nhà chế tạo AI bán những sản phẩm phụ để tự tài trợ cho mình, có một nhóm nhỏ được tài trợ tốt hơn, gắn liền với các viện nghiên cứu nói trên và được DARPA hỗ trợ. Ví dụ, SyNAPSE của MIT mà chúng ta đã thảo luận ở Chương 4, là một thử nghiệm được DARPA tài trợ hoàn toàn nhằm xây dựng một máy tính với bộ não có chức năng và hình dạng tương đương với não của động vật có vú. Bộ não đó sẽ được lắp trước tiên ở những robot được chờ đợi sẽ có trí thông minh cỡ chuột và mèo, và cuối cùng sẽ là các robot hình người. Tám năm qua, SyNAPSE đã tiêu tốn của DARPA 102,6 triệu đô-la. Cũng giống như vậy, NELL của CMU hầu như được DARPA tài trợ, và một số trợ giúp thêm từ Google và Yahoo.

Bây giờ hãy trở lại với Siri. CALO là dự án được DARPA tài trợ để chế tạo Trợ lý (Assistant) Nhận thức (Cognitive), thứ có khả năng Học (Learns) và Tổ chức (Organizes), một dạng Radar O'Reilly◎ kỹ thuật số. Cái tên này

được lấy cảm hứng từ "calonis," tiếng Latin nghĩa là "người phục vụ binh sĩ." CALO được sinh ra tại SRI International, trước đây có tên Viện nghiên cứu Stanford (Stanford Research Institute), một công ty được tạo ra để kinh doanh các sản phẩm phụ từ các dự án mà trường đang nghiên cứu. Mục đích của CALO? Theo như trang web của SRI:

> Mục tiêu của dự án là tạo ra các hệ thống phần mềm nhận thức, nghĩa là các hệ thống có thể suy lý, học từ kinh nghiệm, sai bảo được, giải thích được nó đang làm gì, suy ngẫm từ những kinh nghiệm bản thân, và có phản ứng tốt trước chuyện bất ngờ.[25]

Trong bản thân kiến trúc nhận thức của nó, CALO được cho là sẽ kết hợp các công cụ của AI, bao gồm xử lý ngôn ngữ tự nhiên, máy học, biểu thị kiến thức, tương tác người-máy và đặt kế hoạch mềm dẻo. DARPA tài trợ cho CALO trong thời gian 2003–2008 và trong dự án này có 300 nhà nghiên cứu từ 25 cơ sở, bao gồm Phantom Works của Boeing, các trường Carnegie Mellon, Harvard và Yale. Trong bốn năm, nghiên cứu này đã tạo ra hơn 500 bài báo trong nhiều lĩnh vực liên quan đến AI.[26] Và nó tiêu tốn 150 triệu đô-la tiêri thuế.

Thế nhưng CALO không hoạt động tốt như mong đợi. Tuy vậy, một phần của nó có triển vọng – "cơ chế thực thi" (ngược với cơ chế tìm kiếm) để làm những việc như viết các email và văn bản, thực hiện các tính toán và chuyển đổi, tìm các thông tin chuyến bay và lập ra các nhắc nhở. SRI International, công ty điều phối cả tập đoàn, đã tách Siri thành công ty riêng (từng có tên là *Stealth Company* trong thời gian ngắn) để lấy 25 triệu

đô-la vốn đầu tư phụ trợ, và tiếp tục phát triển "cơ chế thực thi." Vào năm 2008, Apple mua Siri với giá khoảng 200 triệu đô-la.[27]

Hiện nay Siri được tích hợp sâu vào iOS, hệ điều hành của iPhone. Nó là một mảnh nhỏ của thứ mà CALO hứa hẹn sẽ trở thành, nhưng đã thông minh hơn hầu hết các ứng dụng trên điện thoại thông minh khác. Thế còn những binh lính đang chờ đợi CALO? Họ cũng dùng nó – quân đội mang iPhone vào trận chiến, đã cài sẵn Siri và các ứng dụng tác chiến tối mật.[28]

Vì vậy, một lý do *lớn* cho thấy vì sao nguồn tài trợ không phải là một vấn đề đối với AGI và sẽ không làm chậm lại sự bùng nổ trí thông minh, đó là vì chúng ta đang sống trong một thế giới mà những người nộp thuế như bạn và tôi đang phải trả tiền cho việc phát triển AGI, cho từng bộ phận thông minh, qua DARPA (Siri), Hải quân (LIDA), và qua các cơ quan công khai hoặc bí mật khác của chính phủ. Thế rồi chúng ta lại phải trả lần nữa, lần này là cho các chức năng mới quan trọng trong iPhone và máy tính. Thực tế SRI International đã bán một sản phẩm phụ *khác* của CALO có tên Trapit. Nó là một dạng "người trợ giúp nội dung," một công cụ tìm kiếm được cá nhân hóa kiêm khám phá web, thứ làm bạn thấy thú vị và hiển thị chúng ở cùng một chỗ.

Một lý do khác cho việc vì sao kinh tế sẽ không làm chậm lại sự bùng nổ trí thông minh: khi AGI xuất hiện, hoặc thậm chí là mới chỉ gần tới, mọi người sẽ muốn có nó. Tôi nhấn mạnh là mọi người. Goertzel chỉ ra rằng khi các hệ thống trí thông minh cấp độ con người tới, chúng sẽ tạo ra những ảnh hưởng sâu rộng trong nền kinh tế thế giới.[29] Những người tạo ra AGI sẽ nhận được các khoản đầu tư cực lớn để hoàn thiện và thương mại hóa công nghệ này. Độ đa dạng của sản phẩm và dịch vụ mà các tác

nhân thông minh ở mức con người có thể cung cấp sẽ là không thể tin được. Ví dụ như các công việc cổ cồn trắng ở mọi thể loại – có ai là không muốn một đội ngũ thông minh bằng con người làm những việc của con người bằng xương thịt suốt ngày mà không cần nghỉ ngơi và không bao giờ sai lầm? Ví dụ như lập trình máy tính, như Steve Omohundro đã nói ở Chương 5. Con người chúng ta là những lập trình viên tồi, còn trí thông minh máy tính sẽ làm tốt hơn hẳn chúng ta trong việc này (và sẽ sớm sử dụng hiểu biết về lập trình, đó vào bản thân chương trình của nó).

Theo Goertzel, "nếu một AGI hiểu về thiết kế của bản thân nó, nó cũng có thể hiểu và cải tiến các phần mềm máy tính khác, do đó sẽ tạo ra một cuộc cách mạng trong ngành công nghiệp phần mềm. Vì phần lớn giao dịch tài chính trên các thị trường chứng khoán Mỹ hiện tại được các phần mềm giao dịch điều khiển, một công nghệ AGI như thế sẽ dễ dàng và nhanh chóng trở nên không thể thiếu trong ngành tài chính. Quân đội và các cơ quan tình báo cũng rất dễ tìm thấy nhiều ứng dụng thực tiễn của một công nghệ như vậy. Cuộc chạy đua điên cuồng này sẽ diễn ra chi tiết thế nào còn là điều tranh cãi, nhưng ít ra chúng ta có thể chắc chắn rằng bất cứ giới hạn nào về tốc độ tăng trưởng kinh tế và môi trường đầu tư trong giai đoạn phát triển AGI sẽ trở nên không quan trọng."[30]

Tiếp theo, robot hóa AGI – cho nó một cơ thể – và nhiều chân trời mới sẽ mở ra. Ví dụ những công việc nguy hiểm như khai thác mỏ, khám phá đại dương và vũ trụ, chiến tranh, hành pháp, chữa cháy. Những dịch vụ như chăm sóc người già và trẻ nhỏ, người hầu, giúp việc, trợ lý cá nhân. Rồi sẽ có robot làm vườn, lái xe, vệ sĩ và huấn luyện viên thể chất. Khoa học, y dược và công nghệ – có ngành nghề nào mà không tiến bộ vượt trội khi có

những đội ngũ nhân viên thông minh như con người, làm việc không mệt mỏi và thay thế lúc nào cũng được?

Kế tiếp, như chúng ta đã thảo luận lúc trước, cạnh tranh toàn cầu sẽ khiến nhiều quốc gia đấu giá cho công nghệ này, hoặc làm cho họ có cách nhìn khác về các dự án AGI trong nước. Goertzel nói, "Nếu một mẫu AGI thử nghiệm hoạt động được đang tiến gần đến trình độ mà ở đó sự bùng nổ trí thông minh có thể xảy ra, các chính quyền trên khắp thế giới sẽ nhận thức được đây là một công nghệ cực kỳ quan trọng, và sẽ cố gắng bằng mọi giá sản xuất cho được một AGI hoàn thiện 'trước các đối thủ.' Thậm chí nền kinh tế cả nước có thể sẽ thanh lọc để hướng tới mục tiêu phát triển được cỗ máy siêu thông minh đầu tiên. Ngược lại với chuyện giới hạn sự bùng nổ trí thông minh, tốc độ tăng trưởng kinh tế sẽ được định nghĩa bằng nhiều dự án AGI khác nhau diễn ra khắp thế giới."[31]

Nói cách khác, nhiều thứ sẽ thay đổi một khi chúng ta chia sẻ hành tinh này với một trí tuệ thông minh ngưỡng con người, và sau đó nó sẽ thay đổi một lần nữa khi sự bùng nổ trí thông minh kiểu Good diễn ra, và ASI xuất hiện.

Nhưng trước khi xem xét các thay đổi đó và những rào cản quan trọng khác đối với việc phát triển AGI và sự bùng nổ trí thông minh, ta hãy tóm lược lại câu hỏi về rào cản nguồn vốn. Nói một cách đơn giản, nó không phải là rào cản. Việc phát triển AGI không thiếu tiền, vì ba lý do. Thứ nhất, có không ít những dự án AI hẹp sẽ bổ trợ hoặc thậm chí trở thành các bộ phận của các hệ thống AI phổ quát. Thứ hai, có nhiều dự án AGI "không giấu giếm" đang hoạt động và tạo nên những đột phá quan trọng với nhiều nguồn tài trợ khác nhau, chưa kể đến những dự án ẩn danh ngầm. Thứ ba, khi công nghệ AI tiếp cận cấp độ AGI, một dòng lũ vốn sẽ đẩy nó đến

đích. Thực tế, lượng tiền này sẽ lớn đến mức giống như là đầu chuột đuôi voi. Tạm gác lại một số rào cản khác, nền kinh tế thế giới sẽ bị lèo lái bởi sự sáng tạo ra trí tuệ nhân tạo mạnh, và được tiếp sức bởi sự lĩnh hội toàn cầu ngày càng lớn về mọi cách, thứ sẽ thay đổi cuộc đời chúng ta.

Sau đây chúng ta sẽ khảo sát một rào cản quan trọng khác – sự phức tạp của phần mềm. Chúng ta sẽ tìm hiểu xem liệu thử thách trong việc chế tạo ra các kiến trúc phần mềm có trí thông minh cấp độ con người có phải đơn thuần là quá khó để thực hiện không, và liệu mọi chuyện trong tương lai có là một mùa đông AI vĩnh cửu không.

12

Sự Phức Tạp Cuối Cùng

Tại sao chúng ta lại tự tin đến thế khi nói sẽ chế tạo được các máy móc siêu thông minh? Vì những tiến bộ trong khoa học thần kinh đã làm sáng tỏ rằng trí thông minh tuyệt vời của con người có nguồn gốc vật lý, và giờ chúng ta đã biết rằng công nghệ có thể làm mọi điều khả thi về mặt vật lý. Watson của IBM chơi trò *Jeopardy!* giỏi như những nhà vô địch con người là một cột mốc quan trọng, minh họa cho sự phát triển của công nghệ xử lý ngôn ngữ. Watson học ngôn ngữ bằng cách phân tích thống kê một lượng văn bản khổng lồ có trên mạng. Khi máy móc trở nên đủ mạnh để mở rộng phân tích thống kê đó nhằm thống nhất ngôn ngữ với các dữ liệu cảm giác, bạn sẽ thua khi tranh luận với chúng nếu bạn nói rằng chúng không hiểu thế nào là ngôn ngữ.[1]

— **Bill Hibbard**
nhà khoa học AI

Có thật quá xa xôi không khi tin rằng cuối cùng chúng ta sẽ khám phá ra những nguyên tắc để tạo ra trí thông minh và gắn chúng vào một cỗ máy, cũng như chúng ta đã dùng kỹ nghệ đảo ngược để làm ra các máy

> móc đặc biệt hữu dụng mô phỏng những thứ trong tự nhiên như ngựa và ống nhả tơ?◎ Tin mới: bộ não người là một thực thể tự nhiên.[2]
>
> — **Michael Anissimov**
> Giám đốc Truyền thông, MIRI

> **Thành kiến thông thường** – sự từ chối lập kế hoạch hoặc phản ứng lại trước một thảm họa chưa từng xảy ra.[3]

Một số chủ đề thường nảy sinh khi chúng ta khảo sát sự bùng nổ trí thông minh. Theo nhận định của đa số, AGI, khi đạt tới, sẽ là một hệ thống phức tạp, và các hệ thống phức tạp thường gặp sự cố, dù cho trong đó có phần mềm hay không. Các hệ thống AI và kiến trúc nhận thức mà chúng ta đã bắt đầu khảo sát là những dạng hệ thống mà tác giả của *Normal Accidents,* Charles Perrow, đánh giá là phức tạp đến mức chúng ta không thể lường trước được sự đa dạng của sự kết hợp các lỗi hỏng có thể xảy ra. Chẳng phải là quá khi nói rằng AGI rất có thể sẽ được tạo ra trong một kiến trúc nhận thức với kích thước và độ phức tạp lớn hơn mạng đám mây gồm 30.000 vi xử lý của Cycle Computing. Và như công ty đó đã tự khoe, Nekomata là một hệ thống quá phức tạp để con người có thể theo dõi (tức là *hiểu)* nó.

Thêm vào đó, có một thực tế đáng lo ngại là một số bộ phận của hệ thống AGI, chẳng hạn các giải thuật di truyền và các mạng neuron, về cơ bản là không thể hiểu được – chúng ta thực sự không hiểu tại sao chúng lại đưa ra những quyết định này chứ không phải là những quyết định khác. Tuy thế, trong tất cả những người làm việc trong ngành AI và AGI, chỉ một thiểu số là nhận thức được về những hiểm họa nơi chân trời. Đa số họ không lập kế hoạch để đề phòng những kịch bản thảm họa hoặc đề ra

những cơ chế hành động để hạn chế thương vong. Tại Chernobyl và đảo Three Mile, các kỹ sư hạt nhân có hiểu biết sâu sắc về các kịch bản và cách xử lý trong tình trạng khẩn cấp, thế mà họ vẫn thất bại, không thể can thiệp hiệu quả. Vậy những người không chuẩn bị sẽ có bao nhiêu cơ hội để kiểm soát một AGI?

Cuối cùng, hãy xem xét DARPA. Không có DARPA, khoa học máy tính và tất cả những gì chúng ta có từ nó sẽ chỉ ở tình trạng sơ khai. AI sẽ thua kém hiện nay rất nhiều, nếu không muốn nói là chưa ra đời. Nhưng DARPA là một cơ quan phòng vệ. Liệu DARPA có lường trước được độ phức tạp và bất khả tri của AGI? Liệu họ có biết ngoài những mục tiêu ban đầu, AGI sẽ có những động lực riêng của nó không? Liệu DARPA có cho phép biến AI cao cấp thành vũ khí trước khi họ tạo ra được các nguyên tắc đạo đức cho việc sử dụng nó?

Có thể bạn sẽ không thích câu trả lời cho những câu hỏi trên, đặc biệt là khi tương lai nhân loại phụ thuộc vào nó.

• • •

Hãy xét đến rào cản tiếp theo của sự bùng nổ trí thông minh – sự phức tạp của phần mềm. Mệnh đề như sau: chúng ta sẽ không bao giờ đạt tới AGI, hoặc trí thông minh cấp độ con người, bởi vấn đề chế tạo trí thông minh cấp độ con người sẽ cho thấy là quá khó. Nếu chuyện đó xảy ra, không AGI nào có khả năng tự cải tiến đủ tốt để khởi phát sự bùng nổ trí thông minh. Nó sẽ không bao giờ chế tạo được một phiên bản thông minh hơn bản thân nó, dù chỉ một chút, và vòng lặp không bao giờ xảy ra. Những hạn chế tương tự cũng sẽ áp dụng cho giao diện người-máy – chúng sẽ

tăng cường và trợ giúp cho trí thông minh con người, nhưng không bao giờ thực sự vượt trội.

Khi chúng ta tìm hiểu vấn đề sự phức tạp của phần mềm, hãy cùng xem xét luôn chuyện lâu nay con người đã cố gắng vượt qua nó thế nào. Vào năm 1956, John McCarthy, được gọi là "cha đẻ" của trí tuệ nhân tạo (ông đã đặt ra thuật ngữ này), cho rằng toàn bộ vấn đề AGI sẽ được giải quyết trong vòng sáu tháng. Vào năm 1970, nhà tiên phong về AI Marvin Minsky đã nói: "Trong khoảng từ ba đến tám năm, chúng ta sẽ có một cỗ máy với trí thông minh nói chung như một con người bình thường." Nếu xét đến tình trạng khoa học thời đó, và với lợi thế nhìn lại, cả hai đều quá ngạo mạn, theo một nghĩa kinh điển. Ngạo mạn, hay *hubris,* xuất phát từ tiếng Hy Lạp có nghĩa là kiêu ngạo, và thường là kiêu ngạo trước các vị thần. Tội ngạo mạn thường được gán cho người cố gắng làm những việc ngoài giới hạn của con người. Hãy nghĩ đến Icarus đã cố bay, Sisyphus đã đánh lừa thần Zeus (dù chỉ chốc lát), và Prometheus đưa lửa cho con người. Pygmalion theo thần thoại là một nhà điêu khắc, đã đem lòng yêu một trong những bức tượng ông tạo ra, Galatea, tiếng Hy Lạp nghĩa là "tình yêu đang ngủ." Song ông không phải chịu tội. Thay vào đó, Aphrodite, Nữ thần Tình yêu, đã làm cho Galatea trở thành người. Hephaetus, vị thần kỹ thuật của Hy Lạp, thường xuyên chế ra những cỗ máy tự động bằng kim loại giúp ông luyện kim, và đó chỉ là một trong nhiều câu chuyện. Ông tạo ra Pandora *cùng* chiếc hộp của nàng, và Talos người khổng lồ bằng đồng để bảo vệ đảo Cretes khỏi cướp biển.

Paracelsus, nhà giả kim vĩ đại thời trung cổ, nổi tiếng vì đã kết nối dược học với hóa học, nghe đâu đã tinh chỉnh một công thức cho phép tạo ra các sinh vật giống người, và những thứ nửa người nửa vật, có tên homunculi.

Chỉ cần cho xương người, tóc, và tinh dịch vào đầy một cái túi, sau đó chôn xuống hố cùng với một ít phân ngựa. Đợi 40 ngày. Một đứa bé giống người sẽ được sinh ra, và sẽ sống nếu cho nó uống máu. Nó sẽ luôn tí hon nhưng sẽ nghe lời chủ cho đến khi nó phản lại và chạy mất. Nếu bạn muốn pha trộn con người với một loài vật khác, ví dụ như ngựa, hãy thay tóc người bằng tóc ngựa. Tuy nhiên, trong khi tôi có thể nghĩ ra khoảng 10 công dụng của một người tí hon (vệ sinh đường ống lò sưởi, lấy lông chó ra khỏi robot quét nhà Roomba…), thì tôi lại không thể nghĩ ra lợi ích gì cho một con nhân mã tí hon.

Trước khi Phòng nghiên cứu robot của Đại học MIT và nhân vật Frankenstein của nhà văn Mary Shelley tồn tại, đã có một truyền thuyết Do Thái về golem. Như Adam, một golem là một tạo vật giống đực được tạo ra từ đất. Không giống Adam, nó không được Chúa thổi vào sự sống, mà bởi các rabbi (các đạo sĩ Do Thái giáo huyền bí, tin vào một vũ trụ có trật tự và sự thần thánh của con số) niệm câu thần chú gồm các từ ngữ và con số. Tên của Chúa, được viết trên giấy và đặt vào miệng nó, sẽ giữ cho tạo vật câm và liên tục lớn này "sống động." Trong huyền thoại Do Thái, các rabbi có phép thuật đã tạo ra những golem để làm đầy tớ hoặc phục vụ trong nhà. Golem nổi tiếng nhất tên là Yosele, hay Joseph, đã được tạo ra vào thế kỷ 16 bởi trưởng Rabbi Yehuda Loew của thành phố Prague. Vào một thời kỳ mà người Do Thái vẫn bị buộc tội dùng máu của trẻ em Ki-tô để tạo ra thứ bánh mì không men matzoth, Yosele đã bận rộn đi bắt các kẻ thủ ác vô đạo đó, ném lũ tội đồ vào các nhà ngục của Prague, và nói chung đã giúp Rabbi Loew chống lại tội ác. Cuối cùng, theo huyền thoại, Yosele đã hóa điên. Để cứu người Do Thái đồng đạo, Rabbi đã đánh nhau với golem, và lấy đi mảnh giấy thổi sức sống khỏi mồm nó. Yosele trở lại

thành một đống đất sét. Trong một phiên bản khác, Rabbi Loew bị golem khổng lồ đè lên, nghiền nát, một kết cục xứng đáng cho hành động sáng tạo ngạo mạn. Trong một phiên bản khác nữa, vợ Rabbi Loew bảo Yosele đi lấy nước. Nó làm điều đó không ngừng cho đến khi ngôi nhà của người tạo ra nó bị ngập lụt. Trong khoa học máy tính, không biết chương trình của bạn có hành động như thế hay không, gọi là "bài toán dừng." Nghĩa là, một chương trình tốt sẽ chạy cho đến khi nó được lệnh dừng lại, và nói chung không thể biết chắc liệu một chương trình nào đó có dừng lại hay không. Áp dụng vào đây, vợ của Rabbi Loew nên nói rõ cần phải lấy *bao nhiêu* nước, ví dụ như 100 lít, và có lẽ Yosele hẳn sẽ dừng lại khi lấy đủ. Trong câu chuyện này bà đã không làm thế.

Bài toán dừng là một khó khăn thực sự đối với các nhà lập trình, vì sẽ không biết được chương trình của họ có những vòng lặp vô hạn ẩn giấu trong mã như thế nào cho đến khi nó chạy. Và có một thứ thú vị về bài toán dừng, đó là không thể tạo ra một chương trình có khả năng quyết định xem chương trình của bạn *có* lỗi dừng hay không. Bộ phân tích gỡ rối nghe thì có vẻ được, nhưng không ai khác ngoài Alan Turing đã khám phá ra nó vô tác dụng (và ông khám phá ra điều đó trước khi có máy tính *và* lập trình). Ông nói rằng bài toán dừng là không thể giải, bởi nếu một bộ gỡ rối đi đến điểm có lỗi dừng trong chương trình cần gỡ rối, nó cũng sẽ nhảy vào vòng lặp vô hạn trong khi đang phân tích nó, và sẽ không bao giờ xác định được có lỗi dừng hay không. Bạn, người lập trình, sẽ phải đợi nó trả lời với cùng một thời gian như khi bạn đợi chương trình nguyên gốc bị treo. Nghĩa là, một thời gian rất dài, có lẽ là mãi mãi. Marvin Minsky, một trong những cha đẻ của trí tuệ nhân tạo, đã chỉ ra rằng "mọi cỗ máy với trạng thái hữu hạn, nếu cứ để tự chạy hoàn toàn, cuối cùng sẽ rơi vào một kiểu vòng lặp

hoàn hảo. Độ dài của một vòng lặp sẽ không thể lớn hơn số các trạng thái nội tại của cỗ máy đó ” Nghĩa là, một máy tính với bộ nhớ cỡ trung khi chạy một chương trình chứa lỗi dừng sẽ cần một thời gian rất dài để đi hết một vòng lặp, và đó là điều kiện cần để chương trình gỡ rối nhận ra lỗi dừng. Dài bao nhiêu? Dài hơn tuổi của vũ trụ, đối với một số chương trình. Cho nên với các mục tiêu *thực dụng,* thì bài toán dừng nghĩa là sẽ không thể biết được điểm kết thức của bất cứ chương trình nào.

Một khi Rabbi Loew nhận thấy Yosele không có khả năng dừng lại, ông có thể sửa chữa nó với một bản vá (một sự thay đổi mã của nó), trong trường hợp này là lấy mảnh giấy ra khỏi miệng gã khổng lồ, trên đó ghi tên Chúa. Cuối cùng thì Yosele bị cho dừng và lưu trữ, như thần thoại đã nói, tại tầng gác mái của Giáo đường Do Thái Old New tại Prague, và nó sẽ sống lại vào ngày tận thế. Rabbi Loew, một người có thực trong lịch sử, được chôn tại Nghĩa trang Do Thái Prague (đầy thỏa đáng, không xa mộ của Franz Kafka). Huyền thoại về Yosele vẫn rất sống động trong các gia đình có nguồn gốc Do Thái Đống Âu, và thậm chí vào cuối thế kỷ trước trẻ con vẫn được dạy những lời chú sẽ đánh thức golem vào thời điểm tận thế.

Những dấu ấn của Rabbi Loew đã in đậm vào mọi sự kế thừa văn hóa về golem, từ câu chuyện về Frankenstein, đến *Lord of the Rings* (Chúa tể của những chiếc nhẫn) của J.R.R. Tolkien, cho đến máy tính Hal 9000 trong bộ phim kinh điển *2001: A Space Odyssey* (2001: Chuyến du hành không gian) của Stanley Kubrick. Những chuyên gia về khoa học máy tính được Kubrick nhờ tư vấn về robot giết người bao gồm Marvin Minsky và I. J. Good. Lúc đó Good mới chỉ vừa viết về sự bùng nổ trí thông minh, và dự đoán nó sẽ xảy ra sau khoảng hai thập niên. Bởi những chỉ dẫn về Hal

cho Kubrick mà vào năm 1995, Good đã được bổ nhiệm vào Viện hàn lâm Điện ảnh Nghệ thuật và Khoa học (Mỹ), điều có lẽ đã làm ông vừa khó xử vừa thích thú.

Theo lịch sử của Pamela McCorduck, tác giả về AI, những phỏng vấn của bà đã hé lộ một số đông những nhà tiên phong về khoa học máy tính và trí tuệ nhân tạo tin rằng họ là hậu duệ trực tiếp của Rabbi Loew. Trong đó có John von Neumann và Marvin Minsky.

• • •

Song, theo một nghĩa nào đó, với sự trợ giúp từ công nghệ, chúng ta đã vượt qua AGI, hay ngưỡng thông minh của bất kỳ người nào rồi. Chỉ cần kết hợp một người có IQ cỡ trung bình với cơ chế tìm kiếm của Google là bạn đã có một nhóm thông minh hơn con người – một con người mà trí thông minh (intelligence) của anh ta đã được tăng cường (augmented). *IA* thay vì *AI*. Vernor Vinge tin rằng đây là một trong ba con đường chắc chắn sẽ dẫn tới sự bùng nổ trí thông minh trong tương lai, khi một thiết bị được gắn vào não bạn, cung cấp cho nó thêm tốc độ, bộ nhớ và *trí thông minh*.

Trong đầu bạn hãy nghĩ về một người thông minh nhất, và cho anh ta đối đầu với nhóm người giả định của chúng ta — đội Google trong một bài kiểm tra kiến thức và suy luận thông thường. Nhóm người-Google chắc chắn sẽ thắng. Trong việc giải quyết vấn đề khó thì người trí thông minh hơn có khả năng thắng, dù với sự hỗ trợ từ các kiến thức có trên web, nhóm người-Google sẽ không chịu thua dễ dàng.

Kiến thức có giống như trí thông minh? Không, nhưng kiến thức là sự khuếch đại trí thông minh, nếu trí thông minh một phần là khả năng cư xử

khéo léo và mạnh mẽ trong môi trường bạn đang sống. Nhà tiên phong, nhà chế tạo AI Peter Voss đặt giả thiết rằng nếu Aristotle có được nền tảng kiến thức của Einstein, ông cũng sẽ nghĩ ra được thuyết tương đối rộng.[5] Cơ chế tìm kiếm của Google trên thực tế đã làm tăng năng suất lao động lên nhiều lần, đặc biệt là trong những ngành nghề đòi hỏi nghiên cứu và viết lách. Những công việc trước đây cần nhiều thời gian để nghiên cứu – đi đến thư viện để miệt mài với sách và tạp chí ra định kỳ, thực hiện tìm kiếm Lexis/Nexis©, tìm các chuyên gia và viết thư hoặc gọi điện cho họ – giờ đây trở nên nhanh chóng, dễ dàng và rẻ. Tất nhiên, sự tăng năng suất này còn nhờ vào bản thân Internet. Nhưng bạn sẽ bị ngợp trước đại dương thông tin rộng lớn này nếu không có những công cụ thông minh để truy xuất ra một phần nhỏ bạn cần. Google đã làm điều đó thế nào?

Một giải thuật thuộc sở hữu của Google có tên PageRank gán cho mọi trang trên Internet một điểm số từ 0 đến 10. Điểm 1 trên PageRank (đặt theo tên nhà đồng sáng lập Google Larry Page, chứ không phải vì nó xếp hạng trang web) nghĩa là một trang có hai lần "giá trị" so với một PageRank điểm 0. Điểm 2 có nghĩa là hai lần giá trị của điểm 1, v.v…

"Giá trị" thế nào là tùy vào nhiều biến số. Kích cỡ là quan trọng – các trang web lớn hơn sẽ tốt hơn, và các trang lâu đời hơn cũng vậy. Trang web có nhiều nội dung – từ ngữ, hình ảnh, các tùy chọn tải về– sẽ có điểm cao hơn. Trang web có nhanh không, và từ nó có bao nhiêu đường dẫn đến các trang web có chất lượng cao khác? Những yếu tố này và một số yếu tố khác nữa tạo nên thứ hạng trên PageRank.[6]

Khi bạn nhập vào một từ hoặc một câu, Google sẽ thực hiện việc phân tích so khớp siêu văn bản để tìm ra những trang thích hợp nhất với tìm kiếm của bạn. Phân tích so khớp siêu văn bản tìm các từ hoặc câu mà bạn

đã nhập vào, nhưng cũng thăm dò cả nội dung trang web, bao gồm sự phong phú của các mẫu chữ, loại trang web, và các từ đó được đặt ở đâu. Nó nhìn vào việc các từ bạn muốn tìm được trang web này và cả các trang web gần gũi được sử dụng thế nào. Vì PageRank đã chọn ra những trang web quan trọng nhất trên Internet rồi, nên Google không cần phải định giá lại toàn web mà chỉ chú trọng các trang có chất lượng cao nhất. Sự kết hợp của việc so khớp văn bản và xếp hạng trang web cho ra hàng ngàn kết quả trong vài giây, mili giây, nghĩa là bạn vừa gõ thì đã có.

Vậy một đội ngũ nhân viên công nghệ thông tin hiện nay sẽ có năng suất tăng lên bao nhiêu lần so với trước khi có Google? Hai lần? Năm lần? Ảnh hưởng của nó tới nền kinh tế thế nào khi hiệu suất của một lượng lớn người lao động tăng gấp đôi, gấp ba hoặc hơn? Mặt tích cực của nó là chúng ta đạt tổng thu nhập quốc dân cao hơn, nhờ vào tác động của công nghệ thông tin lên năng suất lao động. Mặt tiêu cực là nhiều người lao động sẽ mất việc làm hoặc thay đổi công việc, bởi sự ra đời của một loạt các công nghệ thông tin, trong đó có Google.

Tất nhiên, lập trình một cách khôn ngoan không nên bị nhầm lẫn với trí thông minh, nhưng tôi muốn nói rằng Google và những thứ tương tự *là* những công cụ thông minh chứ không chỉ là các chương trình khôn ngoan. Chúng đã làm chủ được một lĩnh vực hẹp – tìm kiếm – với một khả năng mà con người không thể bì kịp. Hơn thế, Google đặt Internet – kho kiến thức lớn nhất mà loài người từng tạo ra – dưới ngón tay bạn. Và điều đáng nói là tất cả những kiến thức đó sẽ xuất hiện ngay lập tức, nhanh chưa từng thấy (xin lỗi Yahoo, Bing, Altavista, Excite, Dogpile, Hotbot và the Love Calculator◎). Việc viết lách thường được gọi *phơi bày* ký ức của mình. Nó cho phép ta lưu lại những ý nghĩ và ký ức để sau đó thu hồi và phân phát.

Google đã cung cấp những loại trí thông minh mà chúng ta không sở hữu, và không thể phát triển nếu không có nó.

Kết hợp lại, Google và bạn *là* ASI.

Theo một cách tương tự, trí thông minh của chúng ta được mở rộng bởi *sự lưu động* của các công nghệ thông tin mạnh, ví dụ điện thoại di động, nhiều loại có sức mạnh như một máy tính để bàn năm 2000, và mạnh hơn các máy tính dòng chính những năm 1960 một tỉ lần tính trên mỗi đô-la trị giá. Con người chúng ta vốn dĩ di động, và để thực sự thích hợp thì các thiết bị tăng cường trí thông minh của chúng ta cũng phải di động. Internet và các dạng kiến thức khác, chẳng hạn các chương trình chỉ đường, đã có thêm được sức mạnh mới và chiều kích rộng lớn hơn khi chúng ta có thể mang chúng theo khắp mọi nơi. Một ví dụ đơn giản, cái máy tính để bàn của bạn có ý nghĩa gì khi bạn bị lạc trong đêm tối ở một thành phố đầy rẫy tội phạm? Tôi cá rằng nó không giá trị bằng chiếc iPhone đã cài ứng dụng chỉ đường bằng giọng nói.

Vì những lý do như thế, Evan Schwartz, người phụ trách mục *Phê bình Công nghệ* của Đại học MIT đã mạnh dạn đề xuất rằng điện thoại di động đang trở thành "công cụ cơ bản của loài người." Ông lưu ý rằng hơn năm *tỉ* thiết bị đang được sử dụng trên toàn thế giới, nghĩa là gần như mỗi người có một thiết bị.[7] Bước tiếp theo của IA là cho mọi thứ ở điện thoại thông minh vào trong chúng ta – kết nối nó với não chúng ta. Hiện nay chúng ta giao tiếp với máy tính bằng mắt và tai, nhưng trong tương lai, hãy tưởng tượng rằng các thiết bị được cấy vào não sẽ cho phép não kết nối không dây với một mạng đám mây, từ bất cứ đâu. Theo Nicholas Carr, tác giả của *Big Switch* (Công tắc lớn), đó là cái đích mà Larry Page đồng sáng lập Google đang muốn cơ chế tìm kiếm hướng đến trong tương lai.

"Ý tưởng ở đây là bạn không còn cần phải ngồi xuống trước một bàn phím để tìm kiếm thông tin," Carr nói. "Nó trở nên tự động, một kiểu hỗn hợp máy-não. Larry Page đã thảo luận về một kịch bản trong đó bạn chỉ cần nghĩ một câu hỏi, và Google sẽ nói thầm câu trả lời vào tai bạn qua điện thoại di động."[8] Ví dụ bạn hãy xem thông báo gần đây của Google về "Project Glass," chiếc kính cho phép bạn tìm kiếm va hiển thị các kết quả trên Google trong khi bạn đang dạo phố – ngay trong tầm nhìn của bạn.

"Hãy tưởng tượng trong một tương lai rất gần; bạn không quên bất cứ điều gì nữa vì máy tính sẽ nhớ hộ bạn," Eric Schmidt, cựu CEO Google nói. "Bạn sẽ không bao giờ lạc đường. Bạn sẽ không bao giờ cô đơn."[9] Với sự ra đời của một trợ lý ảo thạo việc như Siri trên iPhone, bước đầu tiên của kịch bản đó đã được thực hiện. Trong lĩnh vực tìm kiếm, Siri có một ưu thế khổng lồ so với Google — nó chỉ cung cấp một câu trả lời. Google cung cấp hàng chục ngàn, thậm chí hàng triệu "kết quả," có thể hoặc không phù hợp với tìm kiếm của bạn. Trong một số lĩnh vực giới hạn – tìm kiếm thông thường, tìm đường, tìm các công ty, lập thời biểu, gửi thư điện tử, nhắn tin và cập nhật profile trên mạng xã hội – Siri sẽ cố xác định nội dung và ý nghĩa của cái bạn muốn tìm, để cho bạn một câu trả lời tốt nhất. Đó là chưa kể Siri *nghe* bạn nói, thêm nhận dạng giọng nói vào tìm kiếm di động cao cấp. Nó nói ra câu trả lời. Và nó *học hỏi* công khai. Theo những bằng sáng chế gần đây của Apple, Siri sẽ sớm tương tác với các đại lý trên mạng để mua các sản phẩm như sách và quần áo, thậm chí tham gia vào các diễn đàn online và cuộc gọi trợ giúp khách hàng.[10]

Có thể bạn chưa biết, nhưng chúng ta đã vừa băng qua một cột mốc *khổng lồ* trên con đường tiến hóa của mình. Chúng ta đang trò chuyện với máy móc. Đây là một sự thay đổi lớn hơn GUI rất nhiều. Graphical User

Interface hay Giao diện đồ họa người dùng, do DARPA chế tạo và Apple bán cho người dùng (với đóng góp của Trung tâm nghiên cứu Palo Alto của Xerox, PARC). GUI mang theo triển vọng và ẩn dụ "để bàn" của nó, nghĩa là máy tính để bàn sẽ hoạt động như con người với bàn làm việc và các tệp tin, còn con chuột là thứ sẽ thay thế cho bàn tay. Ý tưởng của hệ điều hành DOS© thì ngược lại – để làm việc với máy tính bạn phải học ngôn ngữ của nó, gồm các câu lệnh cứng nhắc được nhập bằng tay. Hiện giờ chúng ta đang ở một nơi hoàn toàn khác. Các công nghệ của tương lai sẽ thành công hoặc thất bại tùy thuộc vào khả năng học hỏi điều chúng ta làm, và giúp chúng ta làm điều đó.

Cũng như với GUI, các hệ điều hành sẽ phải hoặc đi theo phát minh đột phá của Apple, Siri, hoặc tàn lụi. Và tất nhiên, công nghệ ngôn ngữ tự nhiên sẽ di cư đến máy tính để bàn và máy tính bảng, rồi không lâu sau sẽ đến mọi thiết bị kỹ thuật số, bao gồm lò nướng, máy rửa bát, hệ thống sưởi, điều hòa, hệ thống giải trí và ô tô. Hoặc có lẽ tất cả chúng sẽ được điều khiển bởi chiếc điện thoại trong túi bạn, đã tiến hóa thành một thứ hoàn toàn khác. Nó sẽ không còn là trợ lý ảo, mà là một trợ lý *thực sự,* với các khả năng được nhân lên cùng tốc độ xử lý cao. Và gần như tình cờ, nó sẽ khởi đầu một cuộc đối thoại thực sự giữa người và máy, thứ sẽ còn tồn tại cho đến ngày cuối cùng của loài người.

Nhưng hãy trở lại với hiện tại một lát và lắng nghe Andrew Rubin, Phó Chủ tịch cao cấp về điện thoại di động của Google. Nếu ông được làm theo cách của mình, hệ điều hành Android của Google sẽ không tham gia vào cuộc chơi trợ lý ảo. "Tôi không tin rằng điện thoại của bạn nên là một trợ lý," Rubin nói, và rõ ràng đó là một tuyên bố chiến lược sai lầm nhất mà bạn từng nghe. "Điện thoại của bạn là một công cụ để giao tiếp. Bạn không

nên giao tiếp với điện thoại; bạn nên giao tiếp với ai đó trên điện thoại phía bên kia."[11] Có lẽ ai đó nên nhẹ nhàng nói với Rubin về việc đội của ông đã ngầm đặt chức năng Voice Action vào hệ thống Android rồi. Họ biết tương lai phụ thuộc tất cả vào sự giao tiếp với điện thoại.

• • •

Mặc dù bạn cộng với Google cho ra một loại trí thông minh vượt ngưỡng con người, nhưng nó không phải là loại ra đời từ sự bùng nổ trí thông minh, và nó cũng không dẫn đến đó. Xin nhớ rằng sự bùng nổ trí thông minh đòi hỏi một hệ thống có hai thuộc tính là tự nhận thức và tự cải tiến, mang những siêu sức mạnh cần thiết của máy tính – chạy 24/7 với độ tập trung hoàn toàn, giải quyết các vấn đề với khả năng tự nhân bản, suy nghĩ chiến lược trong chớp mắt… Có thể nói bạn và Google hợp lại tạo nên một dạng thức siêu thông minh đặc biệt, nhưng sự tăng trưởng của nó bị giới hạn bởi bạn và Google. Bạn không thể cung cấp yêu cầu cho Google kiểu 24/7, và Google tuy tiết kiệm thời gian tìm kiếm của bạn, lại lãng phí thời gian của bạn khi buộc bạn phải nhặt ra đáp án tốt nhất trong quá nhiều câu trả lời. Và kể cả có làm việc cùng nhau, nhiều khả năng bạn không phải là một lập trình viên, còn Google thì không thể lập trình được. Cho nên ngay cả khi bạn nhìn thấy những lỗ hổng trong hợp thể này, các nỗ lực để vá chúng của bạn cũng không đủ tốt để tạo ra các tiến bộ nhất định và liên tục. Sự bùng nổ trí thông minh sẽ không xảy ra.

Sự tăng cường trí thông minh (IA) có thể tạo nên một sự bùng nổ trí thông minh không? Tất nhiên, với cùng một tiến trình như AGI. Hãy tưởng tượng một con người, một nhà lập trình có trí thông minh, được sự trợ giúp mạnh mẽ này khiến những kỹ năng lập trình vốn đã giỏi của anh ta trở nên

tốt hơn, nhanh hơn, phong phú hơn và hòa hợp hơn. Kẻ siêu việt giả định này sẽ có thể lập trình ra phiên bản nâng cao tiếp theo của IA.

• • •

Quay lại với sự phức tạp của phần mềm. Mọi biểu hiện đều cho thấy các nhà nghiên cứu máy tính trên thế giới đang làm việc cật lực để pha trộn những thành phần dễ cháy sẽ kích hoạt sự bùng nổ trí thông minh. Liệu sự phức tạp của phần mềm có phải là một rào cản vĩnh viễn đối với thành công của họ?

Có thể cảm nhận được vấn đề phức tạp của phần mềm AGI khó đến thế nào khi trưng câu ý kiến của các chuyên gia về chuyện khi nào thì AGI ra đời. Ở một thái cực, Peter Novig, Giám đốc nghiên cứu của Google, người mà chúng ta đã biết, không quan tâm đến việc đó và chỉ nói AGI hiện vẫn quá xa vời. Trong khi đó, nếu thông tin là chính xác, thì các đồng nghiệp của ông do Ray Kurzweil dẫn đầu đang tiếp tục phát triển nó.

Ở thái cực khác, Ben Goertzel, người cũng như Good tin rằng đạt tới AGI chỉ là vấn đề tiền bạc, tuyên bố rằng trước năm 2020 không phải là thời điểm quá sớm. Ray Kurzweil, người tiên tri công nghệ có lẽ là giỏi nhất từ trước tới nay, tiên đoán sẽ có AGI vào năm 2029, nhưng ASI thì phải chờ đến tận năm 2045. Ông thừa nhận các mối nguy, nhưng tập trung sức mình để củng cố giả thuyết về một chuyến du ngoạn dài êm ả trên đường sinh kỹ thuật số.

Cuộc thăm dò ý kiến không chính thức của tôi với khoảng 200 nhà khoa học máy tính tại Hội thảo AGI gần đây xác nhận điều tôi chờ đợi. Hội thảo AGI thường niên, do Goertzel tổ chức, là cuộc gặp mặt dài ba ngày của

những người đang hoạt động trong ngành trí tuệ nhân tạo phổ quát, hoặc những người chỉ đơn thuần là rất quan tâm như tôi. Họ trưng bày các bài báo, phần mềm thử nghiệm và tranh nhau khoe thành quả. Tôi dự một lần hội thảo được tổ chức tại trụ sở chính của Google ở Mountain View, California, thường gọi là Googleplex. Tôi hỏi những người đến dự khi nào trí tuệ nhân tạo phổ quát sẽ ra đời, và cho họ bốn lựa chọn – 2030, 2050, 2100, không bao giờ. Kết quả như sau: 42% tin rằng AGI sẽ có vào năm 2030; 25% vào 2050; 20% vào 2100; 10% sau 2100 và 2% không bao giờ. Cuộc thăm dò trong nhóm tự chọn này xác nhận cái nhìn lạc quan và những mốc thời gian thu được trong các cuộc trưng cầu ý kiến chính thức hơn, như tôi đã dẫn ở chương 2. Trong phần ý kiến phản hồi, tôi bị trách là đã không cho lựa chọn *trước* năm 2030 vào. Tôi đoán rằng có lẽ khoảng 2% số người muốn chọn thời điểm đạt tới AGI vào năm 2020, và 2% khác thậm chí nghĩ rằng sớm hơn. Tôi vốn thường kinh ngạc khi biết về xu hướng lạc quan này, nhưng giờ thì không. Tôi đã nghe theo lời khuyên của Kurzweil, và nghĩ về sự tiến bộ của công nghệ thông tin theo kiểu hàm mũ, chứ không theo kiểu tuần tự nữa.

Nếu bạn muốn làm sôi động bầu không khí khi bạn đang ở trong một khán phòng đầy những người chuyên sâu về nghiên cứu AGI, hãy tuyên bố "AGI *sẽ không bao giờ* ra đời! Đơn giản là nó quá khó." Người ta, chẳng hạn như Goertzel, sẽ phản ứng với câu này bằng cách nhìn tôi như thể tôi sắp thuyết giảng về Thiết kế thông minh.[◎] Từng là một giáo sư toán học, cũng như Vinge, Goertzel tiên đoán tương lai của AI dựa theo các bài học từ lịch sử của toán giải tích.

"Nếu bạn nhìn vào cách các nhà toán học làm toán giải tích trước Isaac Newton và Gottfried Leibnitz, sẽ thấy họ đã mất cả trăm trang giấy để tính

đạo hàm của một đa thức bậc ba. Họ tính bằng cách vẽ các tam giác, các tam giác đồng dạng và các biểu đồ kỳ quặc. Rất khổ sở. Nhưng giờ thì chúng ta đã có một lý thuyết giải tích hoàn chỉnh và bất kỳ một kẻ ngốc nào ở trường trung học cũng có thể tính đạo hàm của một đa thức bậc ba. Rất dễ."

Cũng như phép giải tích vài thế kỷ trước, nghiên cứu AI sẽ tăng tiến tuần tự cho đến khi quá trình đó dẫn tới sự khám phá ra những quy tắc lý thuyết mới, cho phép nhà nghiên cứu AI cô đọng và trừu tượng hóa nhiều công việc của họ, để rồi tiến trình tới AGI sẽ trở nên đơn giản hơn và nhanh chóng hơn.

"Newton và Leibnitz phát triển những công cụ như đạo hàm của tổng, của tích, của chuỗi, tất cả những quy tắc cơ bản đó bạn đã học trong Giải tích cơ bản," anh nói tiếp. "Trước khi bạn có những quy tắc đó, bạn phải làm mọi phép tính giải tích từ đầu, và như thế thì khó hơn nhiều. Nếu AI là toán học thì chúng ta đang ở trình độ giải tích trước Newton và Leibnitz – nên ngay cả việc chứng minh những thứ cực đơn giản về AI cũng mất hàng núi công sức tính toán cầu kỳ. Nhưng cuối cùng thì chúng ta sẽ có một lý thuyết tốt về trí thông minh, cũng như hiện giờ chúng ta đã có một lý thuyết tốt về giải tích."

Nhưng không có một lý thuyết tốt cũng chẳng sao.

Goertzel nói: "Có lẽ chúng ta cần một sự đột phá khoa học trong lý thuyết khắc nghiệt về trí tuệ nhân tạo phổ quát trước khi chúng ta có thể xây dựng một hệ thống AGI cao cấp. Nhưng hiện tại tôi ngờ rằng điều đó là không cần thiết. Quan điểm bây giờ của tôi là nếu chúng ta cứ tiến từng bước một từ trình độ kiến thức hiện nay – nghĩa là chỉ làm kỹ thuật mà

không cần một nhận thức thông suốt về trí tuệ nhân tạo phổ quát – thì cuối cùng ta cũng sẽ chế tạo được một hệ thống AGI mạnh."[12] Như chúng ta đã thảo luận, dự án OpenCog của Goertzel tổ hợp các phần mềm và phần cứng thành một "kiến trúc nhận thức" mô phỏng hoạt động của ý thức. Kiến trúc này có thể trở nên mạnh mẽ và có lẽ là thứ không thể đoán trước. Trên con đường phát triển của nó, trước khi một lý thuyết hoàn chỉnh về trí tuệ nhân tạo phổ quát ra đời, Goertzel cho rằng OpenCog sẽ đạt tới AGI.

Nghe có vẻ điên rồ? Tạp chí *New Scientist* đề xuất rằng LIDA của Đại học Memphis, một hệ thống giống OpenCog mà ta đã thảo luận ở Chương 11, đã biểu hiện một số dấu hiệu của ý thức sơ đẳng. Nói chung, nguyên tắc chi phối LIDA, được gọi là Lý thuyết Không gian hoạt động Toàn thể, cho rằng tri giác của con người được nuôi dưỡng bởi các giác quan sẽ ngấm vào vô thức cho đến khi nó trở nên đủ quan trọng để được truyền đi khắp bộ não. Đó là ý thức, và nó có thể được đo lường bằng những nhiệm vụ nhận thức đơn giản, vi dụ như ấn một cái nút khi ánh sáng chuyển sang màu xanh lá cây. Dù dùng một cái nút "ảo," nhưng LIDA đạt số điểm tương đương với con người khi được kiểm tra trong các nhiệm vụ này.[13]

Với những công nghệ như thế, cách tiếp cận chờ-và-xem của Goertzel nghe có vẻ rủi ro với tôi. Nó ám chỉ đến việc tạo ra một thứ tôi đã mô tả – trí thông minh máy móc mạnh ngang với con người, nhưng không giống con người và rất khó hiểu. Nó gợi đến những bất ngờ, ví như một ngày một AGI bỗng xuất hiện làm chúng ta không kịp chuẩn bị để đối phó với các tai nạn "thông thường" và tất nhiên là không có những biện pháp an ninh chính thức như AI thân thiện. Nó cũng giống như nói, "Nếu chúng ta đi bộ đủ lâu trong rừng chúng ta sẽ tìm thấy các con gấu đói." Eliezer

Yudkowsky có những lo lắng tương tự. Và giống như Goertzel, ông không nghĩ rằng sự phức tạp của phần mềm sẽ là thứ cản đường.

"AGI là một vấn đề mà bộ não chính là lời giải," ông nói với tôi. "Bộ não người có thể hoạt động – vậy thì nó cũng không phức tạp đến thế. Chọn lọc tự nhiên là thứ ngu ngốc. Nếu chọn lọc tự nhiên có thể giải quyết vấn đề AGI, thì nó cũng không khó đến thế, theo một nghĩa tuyệt đối. Quá trình tiến hóa sinh ra AGI một cách dễ dàng bằng cách thay đổi mọi thứ ngẫu nhiên và giữ lại thứ gì hoạt động tốt. Nó đi theo một tiến trình tuần tự mà không hề có tầm nhìn."

Sự lạc quan của Yudkowsky về việc đạt được AGI bắt đầu với ý tưởng là trí thông minh cấp độ con người đã được chúng ta đạt tới một lần theo kiểu tự nhiên. Con người và tinh tinh có cùng một tổ tiên vào khoảng năm triệu năm trước. Ngày nay não người to gấp bốn lần não tinh tinh. Vậy trong khoảng năm triệu năm, sự chọn lọc tự nhiên "ngu ngốc" đã dẫn tới sự tăng trưởng tuần tự của kích thước bộ não, và tạo ra một động vật thông minh hơn hẳn các loài khác.

Với sự tập trung và tầm nhìn xa, con người "tinh khôn" sẽ có khả năng tạo ra trí thông minh ở cấp độ con người nhanh hơn nhiều so với chọn lọc tự nhiên.

Nhưng một lần nữa, như Yudkowsky đã dẫn, có một vấn đề khổng lồ là nếu có ai đó đạt tới AGI trước khi ông hoặc các nhà nghiên cứu khác tìm ra AI thân thiện hoặc có cách nào đó để kiểm soát tốt AGI. Nếu AGI đến từ cách tiếp cận kỹ thuật tuần tự, từ điểm giao tình cờ của nỗ lực và tai nạn, như Goertzel đề xuất, phải chăng sự bùng nổ trí thông minh sẽ dễ xảy ra? Nếu AGI tự nhận thức và tự cải tiến, như chúng ta đã định nghĩa nó, phải

chăng nó sẽ gắng sức để thỏa mãn các động lực cơ bản mà có lẽ sẽ không tương thích với sự tồn tại của loài người, như chúng ta đã thảo luận ở Chương 5 và 6? Nói cách khác, phải chăng AGI không bị ràng buộc sẽ giết tất cả chúng ta?

"AGI là quả bom nổ chậm," Yudkowsky nói, "là hạn chót buộc chúng ta phải chế tạo được AI thân thiện, dù rất khó. Chúng ta cần AI thân thiện. Ngoại trừ trường hợp công nghệ nano tiêu hủy thế giới, không thảm họa nào trên đời có thể sánh với AGI."

Tất nhiên, sẽ có căng thẳng giữa các nhà lý thuyết AI như Yudkowsky và nhà chế tạo AI như Goertzel. Trong khi Yudkowsky cho rằng tạo ra AGI là một sai lầm kinh khủng trừ phi nó được chứng minh là thân thiện, thì Goertzel muốn phát triển AGI nhanh nhất có thể, trước khi cơ sở hạ tầng trở nên hoàn toàn tự động khiến AGI dễ dàng đoạt lấy quyền lực. Goertzel đã nhận được một số email, dù không phải từ Yudkowsky hay các đồng nghiệp, cảnh cáo rằng nếu anh tiếp tục phát triển thứ AGI không được chứng minh là an toàn, anh đang "phạm tội diệt chủng."[14]

Nhưng ở đây có một nghịch lý. Nếu Goertzel từ bỏ AGI và cống hiến đời mình vào việc khuyên răn những người khác dừng lại, thì không có gì thay đổi cả. Các công ty, chính phủ và trường đại học vẫn sẽ không ngừng tiến tới trong việc nghiên cứu. Chính vì lý do này mà Vinge, Kurzweil, Omohundro và những người khác tin rằng từ bỏ, hay thôi không theo đuổi AGI nữa chẳng phải là một lựa chọn tốt. Thực tế, với sự tồn tại của quá nhiều quốc gia liều lĩnh và nguy hiểm trên thế giới, sự từ bỏ chỉ đơn giản là phó mặc tương lai cho lũ điên rồ và tội phạm.

Một chiến lược phòng ngự có khả năng làm tăng cơ hội sống sót của chúng ta là cách mà Omohundro đã bắt đầu: một khoa học hoàn chỉnh để hiểu và kiểm soát các hệ thống tự nhận thức, tự cải tiến như AGI và ASI. Và bởi thách thức của việc phải phát triển toa thuốc AI thân thiện trước khi AGI xuất hiện, nên việc phát triển khoa học đó phải diễn ra cùng lúc. Rồi khi AGI ra đời, hệ thống kiểm soát nó đã có rồi. Không may cho chúng ta, các nhà nghiên cứu AGI đang đi trước rất xa, và như Vernor Vinge nói, cơn gió kinh tế toàn cầu sẽ thổi căng những cánh buồm của họ.

• • •

Nếu vấn đề phần mềm quả thật là quá khó, thì vẫn còn ít nhất hai con đường cho những người muốn tìm AGI. Chúng là, thứ nhất, giải quyết vấn đề bằng các máy tính mạnh hơn, và thứ hai, dùng kỹ nghệ đảo ngược bộ não.

Việc chuyển thể một hệ thống AI sang AGI bằng sức mạnh phần cứng nghĩa là tăng chức năng hoạt động của phần cứng AI lên, đặc biệt là tốc độ. Trí thông minh và sự sáng tạo sẽ tăng lên khi chúng hoạt động với tốc độ lớn hơn nhiều lần. Để hiểu điều đó, hãy tưởng tượng một người mà chỉ trong *một* phút suy nghĩ được tương đương 1.000 phút. Theo những cách quan trọng, anh ta thông minh hơn nhiều lần so với những người có cùng IQ nhưng suy nghĩ với tốc độ thông thường.[15] Thế nhưng liệu trí thông minh có cần phải bắt đầu ở cấp độ con người để cho sự tăng tốc có ảnh hưởng đến nó? Ví dụ, nếu bạn tăng tốc độ não của chó lên 1.000 lần, bạn sẽ thấy hành vi của nó giống như tinh tinh, hay bạn chỉ có một con chó rất thông minh? Chúng ta biết rằng từ tinh tinh đến con người, *kích thước* não tăng gấp bốn lần, con người có thêm ít nhất một siêu sức mạnh mới –

tiếng nói. Các bộ não lớn hơn tiến hóa tuần tự, chậm hơn nhiều so với tốc độ tăng đều đặn của vi xử lý.

Tóm lại, không rõ là nếu thiếu đi các phần mềm thông minh, thì tốc độ xử lý có bù đắp được, có dẫn tới AGI, và xa hơn là sự bùng nổ trí thông minh hay không. Nhưng dường như chuyện này không phải là không thể.

• • •

Giờ hãy quay lại với cái gọi là "kỹ nghệ đảo ngược" bộ não và tìm hiểu xem tại sao nó có thể là thứ dự phòng cho vấn đề phức tạp của phần mềm. Cho đến nay chúng ta đã nhìn thoáng qua về cách tiếp cận đối lập – chế tạo các kiến trúc nhận thức mà nhìn chung là mô phỏng bộ não tại những khu vực như nhận dạng và điều hướng. Các hệ thống nhận thức này lấy cảm hứng từ cách thức bộ não hoạt động, hoặc – và điều này là quan trọng – cách các nhà nghiên cứu *hiểu* về hoạt động của bộ não. Chúng thường được gọi là *de novo* hay các hệ thống "từ đầu," bởi chúng không đặt trên cơ sở các bộ não thật, và chúng bắt đầu từ con số không.

Vấn đề là, các hệ thống lấy ý tưởng từ các mô hình nhận thức rốt cuộc có thể sẽ thất bại trong việc thực hiện những gì bộ não người có thể làm. Trong khi hiện có nhiều triển vọng trong các lĩnh vực như ngôn ngữ tự nhiên, thị giác máy tính, các hệ thống Hỏi-Đáp và công nghệ robot, vẫn có nhiều bất đồng trong hầu hết mọi khía cạnh về nguyên lý và phương pháp luận, những thứ cuối cùng sẽ tạo ra tiến bộ hướng tới AGI. Các ngành trực thuộc cũng như các lý thuyết phổ quát táo bạo nảy sinh từ những thành công bước đầu hay sức mạnh khởi xướng của một cá nhân hoặc một trường đại học, nhưng rồi chúng cũng lụi tàn nhanh chóng. Như Goertzel nói, không có một lý thuyết nào về trí thông minh và phương pháp điện

toán để đạt tới nó được tất cả chấp nhận. Thêm vào đó, có những chức năng của trí tuệ con người mà công nghệ phần mềm hiện nay dường như chưa đủ khả năng tái tạo, bao gồm học hỏi thông thường, giải thích, tự vấn và kiểm soát chú ý.

• • •

Vậy đã có thành tựu thực sự nào trong ngành AI? Có một chuyện cười cũ thế này: một gã say bị mất chìa khóa xe và đang tìm nó dưới ánh đèn đường. Một cảnh sát đến tìm hộ và hỏi: "Chính xác thì anh mất chìa khóa ở đâu?" Gã chỉ xuống góc tối cuối con phố. "Đằng kia," anh ta nói. "Nhưng ở đây sáng hơn."

Tìm kiếm, nhận dạng giọng nói, thị giác máy tính và phân tích thu hút (loại khả năng máy học mà Amazon và Netflix dùng để gợi ý những thứ bạn có thể sẽ thích) là những lĩnh vực AI được coi là thành công nhất. Dù chúng là sản phẩm của hàng thập niên nghiên cứu, nhưng chúng cũng thuộc về các vấn đề dễ nhất, được khám phá *ở nơi sáng hơn*. Các nhà nghiên cứu gọi chúng là "quả thấp dễ hái." Nhưng nếu mục tiêu của bạn là AGI, thì *mọi* ứng dụng và công cụ của AI hẹp sẽ giống như quả thấp dễ hái, và chúng cũng chỉ khiến bạn tiến gần đích hơn chút xíu. Một số nhà nghiên cứu tin rằng các ứng dụng AI hẹp không hề đóng góp vào quá trình tới AGI. Chúng là các ứng dụng tích hợp đặc biệt. Và không hệ thống trí tuệ nhân tạo nào hiện nay có được chút ít mùi vị tương ứng với tính người nói chung. Chắc bạn đang thất vọng vì AI hứa hẹn thì nhiều nhưng không làm được gì? Có hai quan điểm phổ biến khiến bạn phải nghĩ lại.

Thứ nhất, như Nick Bostrom, Giám đốc Viện Tương lai nhân loại thuộc Đại học Oxford, nói "Nhiều công nghệ AI tối tân đã ăn sâu vào các ứng

dụng hằng ngày, thường nó không còn được gọi là AI bởi khi một thứ trở nên đủ hữu dụng và đời thường thì nó không được dán nhãn AI nữa."[16] Trước đây không lâu, AI không hiện diện ở ngân hàng, y học, vận tải, cơ sở hạ tầng thiết yếu và xe cộ. Nhưng ngày nay, nếu đột nhiên bạn xóa bỏ tất cả AI khỏi những ngành trên, bạn không thể vay nợ, không có điện để dùng, ô tô của bạn không chạy, hầu hết tàu và tàu điện ngầm không dừng lại. Sản xuất thuốc sẽ ngừng lại, nước bị cắt, và các máy bay thương mại sẽ rơi từ bầu trời. Các cửa hàng hoa quả sẽ rỗng không, và không thể mua chứng khoán. Tất cả những hệ thống AI đó đã được cài đặt vào lúc nào vậy? Trong 30 năm qua, tên gọi "mùa đông AI" là một thuật ngữ dùng để chỉ sự thờ ơ kéo dài của các nhà đầu tư, sau khi những tiên đoán sớm quá lạc quan về AI sụp đổ. Nhưng *thực sự* không có mùa đông nào cả. Để tránh vết nhơ của cái nhãn "trí tuệ nhân tạo," các nhà khoa học sử dụng những thuật ngữ mang tính kỹ thuật hơn như máy học, tác nhân thông minh, suy luận xác suất, hệ thống neuron cao cấp, và nhiều nữa.

Và vấn đề sự chấp nhận vẫn còn đó. Những lĩnh vực một thời được cho là hoàn toàn thuộc về con người – ví dụ cờ vua và *Jeopardy!* – hiện giờ đã bị máy tính thống trị (dù chúng ta vẫn được phép chơi). Nhưng bạn có coi trò chơi cờ vua trong máy tính để bàn của bạn là "trí tuệ nhân tạo" không? Watson của IBM có giống con người, hay chỉ là một hệ thống Hỏi-Đáp đặc biệt chạy trên siêu máy tính? Chúng ta sẽ gọi các nhà khoa học là gì khi mà máy tính, giống như Golem, một chương trình được Hod Lipsom tại Đại học Cornell đặt cho cái tên khá thích hợp, đã bắt đầu biết nghiên cứu khoa học? Ý tôi là: kể từ ngày John McCarthy◎ đặt cho khoa học về trí thông minh máy móc một cái tên, các nhà nghiên cứu đã luôn phát triển AI một

cách tận tụy và năng nổ, khiến nó đang trở nên thông minh hơn, nhanh hơn và mạnh hơn theo thời gian.

Thành công của AI trong các lĩnh vực như cờ vua, vật lý và xử lý ngôn ngữ tự nhiên kéo theo một quan sát quan trọng thứ hai. Những thứ khó thì dễ, và những thứ dễ thì khó. Châm ngôn này được biết đến với cái tên Nghịch lý Moravec, bởi Hans Moravec, nhà tiên phong trong AI và công nghệ robot, đã diễn đạt nó chính xác nhất trong cuốn sách robot kinh điển của ông, *Mind Children* (Những đứa trẻ ý thức): "Để máy tính làm các bài kiểm tra trí thông minh hoặc chơi cờ đam giỏi như một người trưởng thành thì tương đối dễ, nhưng để nó có được các kỹ năng của đứa trẻ một tuổi như nhận thức và di chuyển thì là một điều khó khăn, thậm chí bất khả."[17]

Các câu đố khó đến nỗi chúng ta không thể không mắc sai lầm, như khi chơi *Jeopardy!* hay áp dụng Định luật Thứ hai về Nhiệt động học của Newton, chịu thua các AI được lập trình tốt làm trong vài giây. Đồng thời, chưa có hệ thống thị giác máy tính nào có thể phân biệt được giữa chó với mèo – điều mà hầu hết các em bé hai tuổi làm được.◎ Trong một chừng mực nào đó, đây là các vấn đề mang tính bất khả so sánh, khi mà nhận thức trình độ cao đối đầu với kỹ năng cảm giác vận động trình độ thấp. Nhưng nó sẽ khiến những người xây dựng AGI khiêm tốn hơn, vì họ khát khao thấu hiểu toàn bộ trí thông minh của con người. Steve Wozniak, người đồng sáng lập Apple, đã đề xuất một phương án "dễ dàng" thay cho bài kiểm tra Turing, để cho thấy sự phức tạp của các nhiệm vụ đơn giản. Chúng ta sẽ chấp nhận một robot là thông minh, Wozniak nói, nếu nó có thể đi bộ vào căn nhà nào đó, tìm máy pha cà phê và các nguyên liệu, rồi pha cho chúng ta một tách cà phê. Bạn có thể gọi cái này là Bài kiểm tra Người pha cà phê.◎ Nhưng có lẽ nó khó hơn bài kiểm tra Turing, bởi nó

bao gồm các công nghệ AI cao cấp như suy lý, vật lý, thị giác máy tính, truy cập vào cơ sở dữ liệu kiến thức khổng lồ, điều khiển chính xác các bộ dẫn động robot, xây dựng một cơ thể robot dùng vào việc hằng ngày, và còn nữa.

Trong một bài báo có nhan đề *The Age of Robots* (Kỷ nguyên Robot), Moravec cung cấp một dấu hiệu cho nghịch lý cùng tên của ông. Tại sao những thứ khó thì dễ, và những thứ dễ thì khó? Bởi bộ não chúng ta đã tập luyện và hoàn thiện những thứ "dễ," bao gồm thị giác, vận động và di chuyển, kể từ khi những tổ tiên chưa phải là người của chúng ta bắt đầu *có* não. Các thứ "khó" như suy luận là những kỹ năng thu được tương đối gần đây. Và thử đoán xem, chúng thật ra dễ hơn, chứ không khó hơn. Kỹ thuật điện toán khiến ta hiểu ra điều đó. Moravec viết:

> Khi nhìn lại thì dường như theo một nghĩa tuyệt đối, suy luận dễ hơn cảm nhận và hành động rất nhiều – một quan điểm không khó để giải thích theo phạm trù tiến hóa. Trong hàng trăm triệu năm, sự tồn tại của con người (và tổ tiên của họ) đã luôn phụ thuộc vào việc nhìn ngắm và di chuyển trong thế giới thực, và để cạnh tranh thì phần lớn bộ não đã được tổ chức để làm nhiệm vụ này với hiệu suất cao. Nhưng chúng ta không trân trọng kỹ năng to lớn này vì nó có ở mọi người và hầu hết các loài động vật – nó là thứ thông thường. Trong khi đó, suy nghĩ duy lý, ví dụ như chơi cờ, là một kỹ năng mới có, có lẽ mới tồn tại chưa đến trăm ngàn năm. Những bộ phận của não người dùng vào công việc này chưa được tổ chức thật tốt, và theo nghĩa tuyệt đối thì chúng ta chưa giỏi lắm. Nhưng đến tận gần đây, chúng ta vẫn không có đối thủ cạnh tranh để thấy được điều đó.[19]

Và đối thủ cạnh tranh đó tất nhiên là máy tính. Để cho một máy tính làm được điều gì đó thông minh, các nhà nghiên cứu phải tự xem xét bản thân và các *Homo sapiens* khác một cách cẩn thận, thăm dò độ nông sâu của trí thông minh chúng ta. Trong điện toán, cẩn thận trọng để mô hình hóa các ý tưởng bằng toán học. Trong lĩnh vực AI, mô hình hóa sẽ hé lộ những quy tắc và tổ chức ẩn sau những gì chúng ta vẫn làm với não của mình.[20] Vậy thì tại sao không đi trước đón đầu và đơn giản là nhìn vào cách bộ não hoạt động từ *bên trong* não, bằng cách xem xét một cách tỉ mỉ các neuron, các sợi trục và các đuôi gai thần kinh? Tại sao không đơn giản là tìm xem mỗi đám neuron trong não đang làm gì, và mô phỏng nó bằng các giải thuật? Hầu hết các nhà nghiên cứu AI đều đồng ý rằng chúng ta có thể giải mã được bí mật hoạt động của bộ não ra sao, vậy tại sao không đơn giản là *xây dựng một bộ não?*

Đó là luận điểm của "kỹ nghệ đảo ngược bộ não," theo đuổi việc tạo ra một mô hình bộ não bằng máy tính rồi dạy cho nó những gì nó cần phải biết. Như chúng ta đã thảo luận, nó có lẽ *chính là* giải pháp để đạt tới AGI nếu sự phức tạp của phần mềm thực sự là quá khó. Nhưng một lần nữa, nếu mô phỏng toàn thể bộ não *cũng* thực sự là quá khó thì sao? Nếu bộ não đang thực sự làm những việc mà chúng ta không thể mô phỏng bằng kỹ thuật? Trong một bài báo gần đây, để phản biện lại cách Kurzweil hiểu về khoa học thần kinh, Paul Alien, người đồng sáng lập Microsoft và Mark Greaves đồng nghiệp của ông viết: "Bộ não người đơn giản là cực kỳ ấn tượng. Mọi cơ cấu đều được hình thành một cách chuẩn xác qua hàng triệu năm tiến hóa để làm một việc nhất định, bất kỳ là việc gì đi nữa… Trong bộ não mỗi cấu trúc đơn lẻ và các mạch neuron đã được hoàn thiện qua quá trình tiến hóa và các yếu tố môi trường."[21] Nói cách khác, 200 triệu năm

tiến hóa đã gọt giũa bộ não thành một công cụ tư duy tối ưu tuyệt vời bất khả bắt chước -

"Không, không, không, không, không, không, không! Tuyệt đối không. Bộ não *không* được tối-ưu hóa, cũng như tất cả các bộ phận khác của động vật có vú."

Đôi mắt của Richard Granger đảo quanh trong cơn hoảng loạn, cứ như tôi đã thả một con dơi vào văn phòng của ông tại Đại học Dartmouth ở Hanover, bang New Hampshire. Dù là một Yankee chính hiệu New England,◎ nhưng Granger trông giống như một ngôi sao nhạc rock với các tố chất của người Anh di cư – cân đối, gương mặt điển trai và mái tóc nâu giờ đã ngả bạc. Ông nồng nhiệt nhưng thận trọng – như một thành viên ban nhạc, luôn hiểu rằng chơi các nhạc cụ điện tử dưới trời mưa là nguy hiểm. Khi còn trẻ, thực ra Granger có tham vọng trở thành ngôi sao nhạc rock, nhưng thay vào đó lại trở thành nhà khoa học thần kinh điện toán tầm cỡ thế giới, có nhiều cuốn sách được xuất bản và hơn 100 bài báo được bình duyệt. Từ trên văn phòng với nhiều cửa sổ nhìn xuống khuôn viên trường, ông điều hành Phòng nghiên cứu kỹ thuật não của Đại học Dartmouth. Chính nơi đây, vào năm 1956, tại Hội thảo nghiên cứu mùa hè của trường Dartmouth về Trí tuệ nhân tạo, AI lần đầu tiên đã được đặt tên. Hiện nay ở Dartmouth, tương lai của AI nằm trong khoa học thần kinh điện toán – ngành nghiên cứu các nguyên lý điện toán mà bộ não sử dụng để thực hiện các công việc.

"Mục tiêu của chúng tôi trong khoa học thần kinh điện toán là hiểu được bộ não đủ để có thể mô phỏng chức năng của nó. Cũng như những robot đơn giản ngày nay đang thay thế những khả năng thể chất của con người trong các công xưởng và bệnh viện, kỹ thuật não sẽ xây dựng những thứ

thay thế cho các kỹ năng tinh thần. Khi đó chúng ta sẽ có thể tạo ra thứ tương tự não, và sửa chữa não người khi chúng bị hư hỏng."[22]

Nếu bạn là một nhà khoa học thần kinh điện toán như Granger, bạn sẽ tin rằng việc mô phỏng bộ não đơn giản chỉ là vấn đề kỹ thuật. Và để tin *điều đó* bạn phải thấy rằng bộ não cao quý của con người, vị vua của tất cả các cơ quan ở động vật có vú, không có gì là quá cao quý. Granger xem xét não theo phạm trù các bộ phận cơ thể người, không có bộ phận nào trong số đó đã tiến hóa đến mức hoàn thiện.

"Hãy nghĩ về nó như thế này." Granger xòe một bàn tay và nhìn chăm chú vào đó. "Chúng ta hoàn toàn không, không, không, không tối ưu khi có năm ngón tay, có tóc xõa xuống mắt, trán thì không có tóc, mũi ở giữa hai mắt thay vì ở hai bên trái phải. Thật nực cười nếu coi bất kỳ điều nào nói trên là tối ưu hóa. *Tất cả* động vật có vú đều có bốn chi, chúng *đều* có mặt, chúng *đều* có mắt ở phía trên mũi và mũi ở phía trên miệng." Và như ta đã biết, chúng ta hầu như đều có bộ não giống nhau. "Tất cả động vật có vú, bao gồm con người, có những tổ hợp khu vực não giống hệt nhau và chúng kết nối với theo cùng một kiểu, đến mức khó tin," Granger nói. Quá trình tiến hóa là thử nghiệm ngẫu nhiên mọi thứ và sau đó chọn lọc chúng, vậy nên *có thể* bạn sẽ nghĩ mọi thứ khác nhau đó đã được kiểm định trong phòng nghiên cứu vĩ đại của tiến hóa và chúng hoặc được chọn hoặc không. Nhưng thật ra chúng không hề được kiểm định."

Tuy thế, quá trình tiến hóa đã tạo ra một thành tựu đáng kể khi nó đem đến bộ não của động vật có vú, Granger nói. Đó là lý do chỉ có vài đột biến nhỏ xảy ra, kể từ các động vật có vú đến chúng ta. Các bộ phận của nó rườm rà, các kết nối trong nó thì không chính xác và chậm, nhưng nó đã sử dụng các nguyên tắc kỹ thuật mà chúng ta có thể học hỏi – những nguyên

tắc phi tiêu chuẩn mà con người chưa hiểu được. Đó là lý do Granger tin rằng việc tạo ra trí thông minh cần phải được bắt đầu với việc nghiên cứu kỹ bộ não. Ông nghĩ rằng các kiến trúc nhận thức *de novo* — những thứ không xuất phát từ những nguyên tắc của não – sẽ chẳng bao giờ thành công.

"Não là cơ quan duy nhất sinh ra ý nghĩ, sự học, nhận thức," ông nói. "Cho đến nay, không có thành tựu kỹ thuật nào bắt kịp những kỹ năng của bộ não trong những việc trên, chứ chưa nói đến vượt qua. Dù đã kỳ công nghiên cứu và được tài trợ dồi dào, nhưng chúng ta chưa có được các hệ thống nhân tạo có khả năng cạnh tranh với con người trong các lĩnh vực nhận dạng gương mặt, xử lý ngôn ngữ tự nhiên và học hỏi từ kinh nghiệm."

Vậy hãy đánh giá đúng tầm quan trọng của bộ não chúng ta. Chính là não, chứ không phải cơ bắp đã khiến chúng ta trở thành giống loài thống trị trên hành tinh. Chúng ta không ngồi trên đỉnh quyền lực bởi chúng ta đẹp hơn các loài động vật đang tranh ăn với chúng ta, hoặc muốn ăn chúng ta. Chúng ta khôn khéo hơn chúng, thậm chí ngay cả khi đó là sự cạnh tranh với các loài người khác. Trí thông minh, chứ không phải cơ bắp, sẽ giành chiến thắng.

Trí thông minh cũng sẽ giành chiến thắng trong một tương lai đang đến rất nhanh, khi con người chúng ta không còn là những thực thể thông minh nhất. Tại sao lại không? Đã khi nào những người có trình độ kỹ thuật thấp kém áp đảo được những kẻ cao cấp hơn chưa? Đã khi nào một giống loài ngu si thắng được một giống loài khác thông minh hơn?◎ Thậm chí, đã khi nào một giống loài thông minh cho phép một giống loài khác chỉ kém thông minh hơn chút xíu sống chung, ngoại trừ trong tư cách thú cưng? Hãy xem con người chúng ta đối xử với những họ hàng gần nhất của mình

thuộc họ Người thế nào – tinh tinh, đười ươi và khỉ đột. Những con không lên bàn nhậu, không vào sở thú hoặc rạp xiếc thì đang đối diện nguy cơ tuyệt chủng, sống lay lắt qua ngày.

Tất nhiên, như Granger đã nói, không hệ thống nhân tạo nào làm tốt hơn con người trong các lĩnh vực nhận dạng gương mặt, học hỏi và ngôn ngữ. Nhưng trong một số ngành hẹp, AI có sức mạnh tuyệt đối, vô song. Hãy nghĩ về một thực thể nắm tất cả những sức mạnh đó trong tay, và tưởng tượng rằng nó thông minh thực sự, trọn vẹn. Nó sẽ cam chịu làm công cụ cho chúng ta trong bao lâu? Sau một vòng dạo trong trụ sở chính của Google, nhà sử học George Dyson đã nói điều này về nơi chốn tồn tại trong tương lai của một thực thể siêu thông minh như thế:

> Suốt 30 năm qua tôi đã luôn tự hỏi, chúng ta chờ đợi một dấu hiệu tồn tại thế nào từ một AI thực sự? Tất nhiên không phải là một sự tiết lộ thẳng thừng, thứ có thể sẽ tạo nên một phong trào đối lập đòi rút dây nguồn. Chuyện tích lũy của cải hoặc trở nên giàu có bất thường sẽ là một dấu hiệu, hoặc là một cơn thèm khát không ngừng các dữ liệu thô, dung lượng bộ nhớ và các hệ thống vi xử lý, hoặc một chuỗi các mưu kế thâm sâu để có được một nguồn cung cấp điện tự động, không gián đoạn. Nhưng tôi ngờ rằng dấu hiệu thực sự sẽ là một đám đông vui vẻ, thỏa mãn, những người được nuôi dưỡng đầy đủ về thể chất và trí tuệ sẽ vây quanh AI. Sẽ không cần phải có những tín đồ chân chính, hoặc quá trình tải các bộ não người vào máy tính, hoặc thứ gì quái gở như vậy: đơn giản chỉ là một sự giao tiếp tuần tự, lịch lãm, rộng khắp và có qua có lại giữa chúng ta và một thứ gì đó đang lớn lên. Cho đến nay, giả thuyết này vẫn chưa thể kiểm chứng được.

Dyson nói tiếp, dẫn lời của nhà văn khoa học viễn tưởng Simon Ings:

Khi máy móc vượt qua chúng ta, chúng đã quá phức tạp Và tinh tế đến mức không kiểm soát nổi, nên chúng đã làm điều đó nhanh chóng, nhẹ nhàng và hữu dụng đến nỗi chỉ có một kẻ điên rồ hoặc một nhà tiên tri mới dám than phiền.[24]

—

13

Bản Chất Không Thể Hiểu Được

> Sẽ là hợp lý khi giả định rằng siêu trí tuệ nhân tạo đầu tiên sẽ có sức mạnh siêu việt, bởi cả hai lý do: nó có khả năng vượt trội trong việc lập kế hoạch, và nó có thể phát triển ra các công nghệ mới. Nhiều khả năng nó sẽ không có đối thủ: nó có thể đạt tới bất cứ kết quả nào nó muốn và bẻ gãy mọi ý đồ ngăn cản sự hiện thực hóa mục tiêu hàng đầu của nó. Nó sẽ tiêu diệt mọi tác nhân khác thuyết phục chúng thay đổi hành vi, hoặc chặn đứng các nỗ lực can thiệp. Ngay cả một "siêu trí tuệ nhân tạo bị xiềng xích," đang chạy trên một máy tính không kết nối và chỉ có thể tương tác với phần còn lại của thế giới qua giao diện văn bản, cũng có thể vượt ngục bằng cách thuyết phục những kẻ canh giữ giải phóng cho nó. Thậm chí đã có vài bằng chứng thực nghiệm sơ bộ cho thấy chuyện sẽ diễn ra như vậy.[1]
>
> — **Nick Bostrom**
> Viện Tương lai nhân loại, Đại học Oxford

Khi mà AI đang tiến lên trong quá nhiều mặt trận, từ Siri đến Watson, từ OpenCog đến LIDA, thì khó có thể tin rằng việc đạt tới AGI sẽ không

thành hiện thực bởi vấn đề là quá khó. Nếu cách tiếp cận bằng khoa học máy tính thất bại, thì kỹ nghệ đảo ngược bộ não sẽ thành công, dù mất nhiều thời gian hơn. Đó là mục tiếu của Rick Granger: hiểu thấu bộ não từ dưới lên, bằng cách tái tạo những cấu trúc cơ bản nhất của não trong các chương trình máy tính. Và ông không đừng được việc chế nhạo những người đang nghiên cứu các nguyên lý nhận thức theo kiểu từ trên xuống, bằng khoa học máy tính.

"Họ đang nghiên cứu hành vi con người và cố gắng tìm hiểu xem họ có thể mô phỏng hành vi đó với một máy tính. Công bằng mà nói, chuyện này hơi giống việc cố hiểu một cái xe ô tô nhưng lại không nhìn vào trong xe. Chúng ta nghĩ rằng chúng ta có thể viết được ra trí thông minh là gì. Chúng ta nghĩ rằng chúng ta có thể viết được ra sự học là gì. Chúng ta nghĩ rằng chúng ta có thể viết được ra các kỹ năng thích ứng là gì. Nhưng lý do duy nhất khiến chúng ta có được bất cứ khái niệm nào về chúng là bởi chúng ta quan sát thấy con người làm những việc 'thông minh.' Vấn đề là chỉ nhìn con người làm những việc đó sẽ không cho ta biết được bất kỳ chi tiết cụ thể nào về điều thực sự đang diễn ra. Câu hỏi phản biện ở đây là: đặc điểm kỹ thuật của sự suy lý và sự học là gì? Không có đặc điểm kỹ thuật nào cả, vậy thì họ đang xây dựng mọi thứ trên cơ sở nào, ngoại trừ việc quan sát đơn thuần?"

Và chúng ta biết mình là những kẻ tự quan sát tồi tệ. "Hết lần này đến lần khác, một lượng lớn các nghiên cứu trong tâm lý học, khoa học thần kinh và khoa học nhận thức cho thấy chúng ta kém cỏi kinh khủng khi phải xem xét nội tâm mình," Granger nói. "Chúng ta không hiểu chút nào về hành vi của bản thân, cũng như những quá trình tạo nên chúng." Granger lưu ý rằng chúng ta cũng yếu kém như thế khi phải đưa ra các quyết định

duy lý, cung cấp chính xác lời khai nhân chứng và nhớ lại những gì vừa xảy ra. Nhưng sự giới hạn trong khả năng quan sát của chúng ta không có nghĩa rằng các khoa học nhận thức dựa trên sự quan sát đều là vớ vẩn. Granger chỉ nghĩ rằng chúng là những công cụ sai lầm khi muốn đột nhập khu vườn của trí thông minh.

"Trong khoa học thần kinh điện toán chúng tôi nói rằng 'tốt, vậy thì não người *thực sự làm* cái gì?' " Granger nói. "Không phải cái ta *nghĩ* rằng nó đang làm, và không phải cái ta *muốn* nó làm. Nó thực sự làm cái gì? Có lẽ những thứ đó, lần đầu tiên sẽ cho ta định nghĩa về trí thông minh, định nghĩa về sự thích nghi, định nghĩa về ngôn ngữ."

Việc suy luận các nguyên tắc điện toán của bộ não bắt đầu khi các nhà khoa học kiểm tra xem những cụm neuron trong não đang làm gì. Neuron là những tế bào có khả năng gửi và nhận các tín hiệu điện hóa học. Những phần quan trọng nhất của chúng là sợi trục (chất xơ kết nối neuron với nhau, thường là nơi gửi tín hiệu), khớp (mối nối nơi tín hiệu giao nhau) và sợi nhánh (thường là nơi nhận tín hiệu). Có khoảng 100 tỉ neuron trong não. Mỗi neuron được kết nối với hàng chục ngàn neuron khác. Sự kết nối chằng chịt này khiến các hoạt động của não diễn ra song song, thay vì theo chuỗi như hầu hết các máy tính. Theo thuật ngữ máy tính thì xử lý theo chuỗi nghĩa là xử lý liên tiếp – thực hiện từng phép tính một. Xử lý song song nghĩa là nhiều dữ liệu được thao tác đồng thời – nhiều lúc hàng chục ngàn, thậm chí hàng triệu tính toán xảy ra cùng lúc.

Thử tưởng tượng trong giây lát về một con phố tấp nập, và nghĩ đến tất cả những dữ liệu đầu vào về màu sắc, âm thanh, mùi vị, nhiệt độ và xúc giác đang cùng lúc nhập vào bộ não của bạn qua mắt, tai, mũi, chân tay và da thịt. Nếu bộ não của bạn không phải là một cơ quan có khả năng xử lý

tất cả những thứ đó đồng thời, lập tức nó sẽ bị quá tải. Thay vào đó, các giác quan của bạn thu thập mọi dữ liệu đó, xử lý nó bằng các neuron trong não, và xuất các dữ liệu đầu ra, tức hành vi, ví dụ đứng trong phần đường đi bộ và tránh va chạm với những khách bộ hành khác.

Tập hợp các neuron hoạt động trong cùng một mạch rất giống với mạch điện. Một mạch điện thì có dòng điện, và dòng điện đó được chạy qua các dây dẫn và các bộ phận đặc biệt như điện trở và diode. Trong quá trình đó, dòng điện sẽ thực hiện các chức năng như thắp sáng bóng đèn hoặc chạy máy cắt cỏ. Nếu bạn tạo ra một danh sách những câu lệnh làm nên chức năng hoặc tính toán đó, bạn sẽ có một chương trình máy tính, hoặc một giải thuật.

Những cụm neuron trong não bạn tạo nên những mạch có chức năng hoạt động như những giải thuật. Và chúng không thắp sáng cái gì cả, nhưng chúng nhận diện các gương mặt, lên kế hoạch đi nghỉ hè và viết ra một câu văn. Tất cả được vận hành song song. Bằng cách nào mà các nhà nghiên cứu biết được điều gì đang diễn ra trong những cụm neuron đó? Nói một cách đơn giản, họ thu thập các dữ liệu giải pháp có độ phân giải cao với các công cụ chụp hình neuron đa dạng, từ những điện cực được cấy trực tiếp vào não của động vật, cho đến những máy móc phức tạp như máy chụp cắt lớp (PET) và máy chụp cộng hưởng từ chức năng (fMRI) cho người. Các máy thăm dò thần kinh trong và ngoài não có thể cho biết từng neuron đơn lẻ đang làm gì, trong khi đánh dấu neuron bằng thuốc nhuộm nhạy cảm điện áp sẽ cho thấy những neuron nào đang được kích hoạt. Từ những kỹ thuật này và một số kỹ thuật khác, xuất hiện những giả thuyết có thể kiểm chứng về các giải thuật chi phối mạch não. Người ta cũng đã bắt đầu xác định chức năng chính xác của một số vùng trong não. Ví dụ, hơn một thập

niên nay, các nhà khoa học thần kinh đã biết công việc nhận dạng gương mặt thuộc về một bộ phận của não được gọi là hồi hình thoi◎.

Vậy vấn đề là gì? Nếu một hệ thống điện toán được bắt nguồn từ bộ não (cách tiếp cận bằng khoa học thần kinh điện toán), nó có hoạt động tốt hơn một hệ thống được tạo ra theo kiểu *de novo* (cách tiếp cận bằng khoa học máy tính)?

Có một loại hệ thống khởi nguồn từ bộ não, đó là mạng neuron nhân tạo, nó hoạt động tốt và lâu đến mức đã trở thành xương sống của ngành AI. Như chúng ta đã thảo luận ở Chương 7, ANN (thứ có thể được tạo ra bằng phần cứng hoặc phần mềm) đã được phát minh vào những năm 1960 để vận hành như neuron. Một trong những lợi ích chủ chốt của nó là có thể dạy nó được. Ví dụ, nếu bạn muốn dạy cho một mạng neuron cách dịch từ tiếng Phap sang tiếng Anh, bạn có thể dạy nó bằng cách nhập các văn bản tiếng Pháp và bản dịch chính xác sang tiếng Anh của chúng vào máy. Cách này gọi là học có kiểm soát. Khi có đủ thí dụ, mạng này sẽ nhận ra được các quy luật kết nối các từ tiếng Pháp với các từ tiếng Anh tương ứng.

Trong não, các khớp thần kinh kết nối các neuron, và sự học xuất hiện trong những kết nối đó. Kết nối của khớp càng mạnh thì ký ức càng sâu sắc. Trong ANN, độ mạnh của một kết nối khớp được gọi là "độ nặng" của nó, và nó được biểu thị dưới dạng một xác suất. Một ANN sẽ gán các độ nặng của khớp với các quy tắc phiên dịch ngoại ngữ mà nó thu được từ việc huấn luyện. Huấn luyện càng nhiều thì nó dịch càng tốt. Trong huấn luyện, ANN sẽ học cách nhận ra các lỗi sai của nó, và điều chỉnh các độ nặng của khớp tùy theo đó. Điều đó có nghĩa là mạng neuron có tính chất tự cải tiến.

Sau quá trình huấn luyện, khi văn bản tiếng Pháp được nhập vào, ANN sẽ căn cứ vào các quy tắc được xác suất hóa mà nó thu được trong quá trình học và xuất ra bản dịch tốt nhất. Về bản chất, ANN đang nhận dạng các kiểu mẫu trong kho dữ liệu. Ngày nay, việc tìm ra các mẫu từ trong một lượng lớn dữ liệu hỗn độn là một trong những công việc sinh lợi nhất của AI.

Bên cạnh dịch thuật và khai thác dữ liệu, ANN hiện được dùng trong các AI trò chơi máy tính, trong việc phân tích thị trường chứng khoán và nhận diện các vật thể trong tranh ảnh. Chúng có trong các Chương trình Nhận dạng Ký tự bằng Thị giác dùng để đọc văn bản được in ra và trong các con chip máy tính dùng để điều khiển tên lửa. ANN thêm phần "thông minh" vào "bom thông minh" Chúng cũng sẽ là thứ có tính quyết định trong hầu hết các kiến trúc AGI.

Cần nhắc lại một số điều quan trọng ở Chương 7 về những mạng neuron thường gặp này. Giống các giải thuật di truyền, ANN là những hệ thống "hộp đen." Tức dữ liệu đầu vào – trong ví dụ này là tiếng Pháp – là rõ ràng. Và dữ liệu đầu ra – ở đây là tiếng Anh – cũng dễ hiểu. Nhưng điều gì đã xảy ra giữa hai đầu đó thl không ai biết. Tất cả những gì nhà lập trình có thể làm là huấn luyện ANN bằng các ví dụ, và cố gắng cải thiện dữ liệu đầu ra. Bởi thứ mà các công cụ trí tuệ nhân tạo dạng "hộp đen" xuất ra không bao giờ có thể dự đoán được, chúng sẽ không bao giờ đáng tin và an toàn.

Dựa trên kết quả thu được, các giải thuật xuất phát từ não của Granger chúng minh rằng cách tốt nhất để tìm kiếm tri thông minh có lẽ là đi theo hình mẫu của tạo hóa – bộ não người – thay vì theo các hệ thống *de novo* của khoa học nhận thức.

Vào năm 2007, các sinh viên của ông tại Đại học Dartmouth đã tạo ra một giải thuật thị giác bắt nguồn từ bộ não, nó nhận dạng các vật thể nhanh hơn các giải thuật truyền thống 140 lần. Nó vượt qua 80.000 giải thuật khác để giành giải thưởng 10.000 đô-la của IBM.

Vào năm 2010, Granger và đồng nghiệp là Ashok Chandrashekar đã tạo ra các giải thuật bắt nguồn từ bộ não cho việc học có kiểm soát. Học có kiểm soát được dùng để dạy máy móc các ký tự hình ảnh và nhận dạng giọng nói, lọc thư rác… Được thiết kế để dùng với các vi xử lý song song, những giải thuật từ bộ não của họ thực hiện các công việc chính xác ngang với các giải thuật theo dãy, nhưng *nhanh hơn 10 lần*. Những giải thuật mới này được tìm ra khi quan sát các cụm, hay các mạng neuron thông thường nhất trong não.

Vào năm 2011, Granger và các đồng nghiệp được cấp bằng phát minh vì đã tạo ra một con chip xử lý song song có thể tái cấu hình dựa trên những giải thuật này Điều đó có nghĩa là một số phần cứng thường gặp nhất trong não có thể được tái sản xuất trong một con chip mấy tính. Khi liên kết các con chip đó với nhau, giống như chương trình SyNAPSE của IBM, bạn sẽ tạo ra một bộ não ảo. Và hiện nay, mỗi một con chip này có thể tăng tốc và tăng cường hiệu suất của các hệ thống được thiết kế để nhận dạng gương mặt trong đám đông, tìm bệ phóng tên lửa trong những bức ảnh vệ tinh, tự động dán nhãn cho bộ sưu tập ảnh kỹ thuật số lộn xộn của bạn, và hàng trăm việc khác. Theo thời gian, các mạch khởi nguồn từ não có thể đem tới khả năng chữa trị cho những bộ não bị tổn thương, bằng cách chế tạo các thiết bị hỗ trợ hoặc thay thế cho phần não yếu. Một ngày nào đó, những con chip xử lý song song mà đội của Granger đã phát minh sẽ thay thế cho các "phần ướt" bị hỏng của bộ não.

Trong khi đó, các phần mềm được suy luận từ não đang dần thâm nhập vào cách thức xử lý điện toán truyền thống. Hạch nền là một bộ phận cổ xưa của não có nguồn gốc từ loài bò sát, gắn với việc điều khiển vận động. Các nhà nghiên cứu nhận ra hạch nền sử dụng những giải thuật dạng học hỏi tăng cường để thu thập các kỹ năng. Đội của Granger đã khám phá ra các mạch trong vỏ não, phần được thêm vào não sau cùng, đã tạo ra những hệ thống thứ bậc về các thực tế *và* tạo ra các mối tương quan giữa các thực tế đó, tương tự như các cơ sở dữ liệu phân cấp. Đó là hai cơ chế khác nhau.

Bây giờ đến phần thú vị: các mạch trong hai phần nói trên của não (hạch nền và vỏ não) được kết nối bằng các mạch khác, kết hợp những hiệu năng của chúng. Trong điện toán cũng có một thứ tương tự. Các hệ thống máy tính học hỏi tăng cường hoạt động bằng cách thử-sai, chúng cần kiểm tra vô vàn khả năng để biết được câu trả lời đúng. Đó là cách thức nguyên thủy mà *chúng ta* dùng hạch nền để học các thói quen, ví dụ như cách đi xe đạp hay đánh một quả bóng chày.

Nhưng con người cũng có một hệ thống phân cấp ở vỏ não, thứ cho phép chúng ta thay vì tìm kiếm một cách mù quáng qua tất cả các khả năng thử-sai, thì sẽ phân loại chúng, sắp xếp thứ bậc cho chúng, và chọn lọc các khả năng đó theo cách thông minh hơn nhiều. Sự kết hợp này hoạt động nhanh hơn hẳn, đưa đến những giải pháp tốt hơn nhiều so với các động vật khác, ví dụ các loài bò sát vốn chỉ dùng mỗi hệ thống thử-sai của hạch nền.

Có lẽ thứ cao cấp nhất mà chúng ta có thể làm với sự kết hợp vỏ não – hạch nền là thực hiện các phép thử-sai *trong não* mà thậm chí không cần phải thử nghiệm chúng trong đời thực. Chúng ta có thể làm điều đó nhiều lần bằng cách chỉ cần nghĩ về nó: mô phỏng tất cả trong đầu mình. Các

giải thuật nhân tạo kết hợp những phương pháp này thực hiện tốt hơn nhiều so với từng phương pháp riêng lẻ. Granger giả định rằng điều đó rất giống với ưu thế được tạo ra khi hai hệ thống trong não kết hợp với nhau.

Granger và các nhà khoa học thần kinh khác cũng đã khám phá ra rằng chỉ có một số ít loại giải thuật thống trị các mạch não. Các hệ thống điện toán cốt lõi giống nhau được sử dụng lặp đi lặp lại trong những hoạt động nhận thức và cảm giác khác nhau, chẳng hạn như nghe hoặc suy luận. Một khi những hoạt động này được tái lập trong phần mềm và phần cứng máy tính, có lẽ chúng có thể được nhân đôi một cách đơn giản để tạo ra các module mô phỏng những phần khác nhau của não. Và việc tái lập các giải thuật, ví dụ như giải thuật nghe, sẽ tạo ra những ứng dụng nhận dạng giọng nói tốt hơn. Thực tế, chuyện này đã diễn ra rồi.

Kurzweil là một trong những nhà phát minh đầu tiên áp dụng các hiểu biết về não vào lập trình. Như chúng ta đã thảo luận, ông cho rằng kỹ nghệ đảo ngược bộ não là con đường triển vọng nhất dẫn tới AGI. Trong một bài luận bảo vệ cho quan điểm này và các tiên đoán của ông về những cột mốc công nghệ, ông viết:

> Về cơ bản, chúng ta đang tìm kiếm các phương pháp lấy cảm hứng từ sinh học để tăng tốc các tiến bộ trong ngành AI, một bộ phận lớn của ngành này đang phát triển mà không có những hiểu biết quan trọng về cách mà bộ não thực hiện những chức năng tương tự. Từ công việc của riêng tôi trong lĩnh vực nhận dạng giọng nói, tôi hiểu rằng công việc của chúng ta sẽ tiến triển nhanh hơn nhiều nếu chúng ta có được những hiểu biết sâu về cách bộ não chuẩn bị và biến đổi thông tin dạng âm thanh.[2]

Trở lại với những năm 1990, công ty Kurzweil Computer Technologies đã có đột phá mới trong nhận dạng giọng nói với những ứng dụng được thiết kế để các bác sĩ kê bệnh án bằng lời. Kurzweil đã bán công ty này, và nó trở thành một trong những tiền thân của Nuance Communications, Inc. Bất cứ khi nào bạn dùng Siri, các giải thuật của Nuance sẽ phụ trách phần nhận dạng giọng nói diệu kỳ của nó. Nhận dạng giọng nói là nghệ thuật dịch âm nói sang văn bản (đừng nhầm với NLP là hiểu nghĩa của văn bản viết). Sau khi Siri dịch câu hỏi của bạn sang dạng văn bản, ba khả năng khác của nó bắt đầu làm việc: triển khai NLP, tìm kiếm một cơ sở dữ liệu kiến thức rộng lớn, và tương tác với các nhà cung cấp dịch vụ tìm kiếm Internet như OpenTable, Movietickets và Wolfram Alpha.

Watson của IBM là một dạng Siri phiên bản khủng, và là một nhà vô địch về NLP. Vào tháng 2/2011, nó sử dụng cả hai loại hệ thống khởi nguồn từ não và hệ thống lấy cảm hứng từ não để có được chiến thắng ngoạn mục trước các đối thủ con người trong trò *Jeopardy!*. Cũng như máy tính vô địch cờ vua Deep Blue, Watson là cách để IBM phô diễn kỹ nghệ điện toán của mình trong quá trình theo đuổi lĩnh vực AI. Trò chơi lâu đời này hứa hẹn là một thử thách khó nhằn bởi tính mở của các dấu hiệu và trò chơi chữ. Người chơi phải hiểu được những cách chơi chữ, cách ví von, những điển tích văn hóa đặc thù, và họ phải đặt các câu trả lời dưới dạng các câu hỏi. Tuy nhiên, nhận dạng ngôn ngữ không phải là chuyên môn của Watson. Nó không có khả năng hiểu lời nói. Và vì nó không thể nhìn hoặc cảm nhận, cũng không đọc được, nên khi thi đội của Watson phải nhập bằng tay các câu hỏi của *Jeopardy!*. Và vì Watson cũng không thể *nghe,* nên các dấu hiệu bằng âm thanh và hình ảnh không được sử dụng.

Nào gượm đã, có thật là Watson đã thắng trò *Jeopardy!* không, hay chỉ là một phiên bản đặc biệt của nó thôi?

Kể từ chiến thắng đó, để Watson hiểu được mọi người nói gì, IBM đã tích hợp vào nó công nghệ nhận dạng giọng nói của Nuance. Và Watson đang đọc hàng terabyte dữ liệu về ngành y. Một trong các mục tiêu của IBM là thu nhỏ Watson lại từ kích thước hiện tại – một phòng đầy những máy chủ – thành kích cỡ một cái tủ lạnh và khiến nó trở thành bác sĩ khám bệnh tốt nhất thế giới. Một ngày nào đó không xa, có lẽ bạn sẽ có cuộc hẹn với một trợ lý ảo, nó sẽ đặt cho bạn rất nhiều câu hỏi và cung cấp cho dược sĩ của bạn một đơn thuốc.[3] Không may là Watson vẫn chưa thể nhìn được và chưa thể nhận biết được các dấu hiệu sức khỏe như mắt sáng, má hồng, hay một vết thương do vừa bị bắn. IBM cũng lên kế hoạch đưa Watson vào điện thoại thông minh như một ứng dụng Hỏi-Đáp tối thượng.

• • •

Những kỹ năng khởi nguồn từ não của Watson đến từ đâu? Phần cứng của nó là một hệ thống song song siêu lớn, sử dụng khoảng 3.000 bộ vi xử lý song song để chạy 180 phần mềm khác nhau được viết chuyên dụng.[4] Xử lý song song là khả năng tuyệt vời nhất của bộ não, các nhà phát triển phần mềm mô phỏng nó rất vất vả. Như Granger đã nói với tôi, các bộ vi xử lý song song và phần mềm được thiết kế cho chúng đã không đạt kỳ vọng. Tại sao? Bởi những chương trình viết cho chúng không giỏi chia nhỏ các nhiệm vụ để giải quyết song song. Nhưng như Watson từng trình diễn, các phần mềm song song được cải tiến đã thay đổi tất cả những điều đó, và phần cứng song song đang theo sát đằng sau. Các con chip song song mới đang được thiết kế để tăng tốc đáng kể phần mềm hiện có.

Watson cho thấy kỹ nghệ song song có thể giải quyết một lượng công việc điện toán khổng lồ với tốc độ tia chớp. Nhưng thành tựu chinh của Watson là ở chỗ nó có thể tự học. Các giải thuật của nó tìm các mối tương quan và hình mẫu trong dữ liệu văn bản mà nhà chế tạo đưa cho nó. Bao nhiêu dữ liệu? Các bộ bách khoa toàn thư, báo chí, tiểu thuyết, từ điển đồng nghĩa, toàn bộ Wikipedia, Kinh Thánh – tổng cộng bằng khoảng tám triệu cuốn sách dày, được nó đọc với tốc độ 500 gigabyte (1.000 cuốn sách dày) mỗi giây. Đáng kể ở chỗ những thứ đó bao gồm *cơ sở* dữ liệu từ vựng, phân loại từ vựng (từ vựng được sắp theo thư mục và phân loại) và từ điển tương quan (mô tả từ vựng và cách chúng liên hệ với nhau). Về cơ bản, đấy là một lượng lớn những hiểu biết thông thường về từ ngữ. Ví dụ: "Trần nhà là phần trên của ngôi nhà, không phải phần dưới, ví dụ tầng hầm, và không phải phần bên, ví dụ bức tường." Câu này cho Watson biết một ít về trần nhà, ngôi nhà, tầng hầm và bức tường, nhưng nó cần phải biết định nghĩa của mỗi thứ đó để câu này có thể hiểu được, và cả định nghĩa của từ "phần" nữa. Và nó sẽ muốn xem những thuật ngữ này được dùng trong các câu thế nào, càng nhiều ví dụ càng tốt. Watson có được tất cả những điều đó.[5]

Trong phần hai của trò chơi *Jeopardy!*, gợi ý này được đưa ra: "Điều khoản này nằm trong một bản giao kèo công đoàn nói rằng lương sẽ tăng hoặc giảm tùy thuộc vào một số tiêu chí như chi phí sinh hoạt." Đầu tiên Watson phân tích cú pháp, nghĩa là nó chọn và phân tích những từ chủ chốt của câu văn. Sau đó nó tìm thấy trong nguồn dữ liệu đã được tiếp nhận rằng lương là một thứ có thể tăng hoặc giảm, một bản giao kèo thì có chứa các thuật ngữ về lương, và những bản giao kèo bao gồm các điều khoản. Nó có một gợi ý khác rất quan trọng: tiêu đề của phạm trù đang bàn luận là

“Luật ‘E.’ ” Điều đó cho Watson biết rằng câu trả lời có liên quan đến một thuật ngữ luật thông thường, và nó sẽ bắt đầu với chữ cái “E.” Watson đã thắng các đối thủ người với câu trả lời: “Có phải là điều khoản điều chỉnh trượt giá© không?” Tất cả những việc đó mất ba giây.

Sau khi Watson có được câu trả lời đúng trong phạm trù này, nó trở nên tự tin hơn (và chơi mạnh dạn hơn) khi “nhận ra” nó đang diễn dịch hạng mục này một cách chuẩn xác. Nó thích nghi với trò chơi, hay học cách chơi tốt hơn, trong khi trò chơi đang diễn ra.

Hãy tạm không nghĩ đến *Jeopardy!* một lát và tưởng tượng kỹ nghệ máy học có tính thích nghi và tốc độ cao có thể được ứng dụng để lái xe, lái tàu chở dầu hoặc thăm dò mỏ vàng như thế nào. Hãy nghĩ về toàn bộ sức mạnh có được ở một trí thông minh ngang với con người.

Watson cũng đã biểu thị một loại trí thông minh khá thú vị khác. Phần mềm DeepQA của nó tạo ra hàng trăm câu trả lời tiềm năng và thu thập hàng trăm bằng chứng cho mỗi câu trả lời đó. Sau đó nó lọc và xếp hạng các câu trả lời theo mức độ tin cậy của mỗi câu. Nếu nó không cảm thấy đáng tin vào một câu trả lời, nó sẽ im lặng, bởi trong trò *Jeopardy!* trả lời sai sẽ bị phạt. Nói cách khác, Watson biết nó không biết cái gì. Giờ đây, bạn có thể không tin rằng một phép tính xác suất cấu thành nên sự tự nhận thức, nhưng phải chăng nó là một trong những bước đầu tiên dẫn tới đích? Watson có thực sự *biết* điều gì không?

Vậy nếu các mạch của não được điều khiển bởi các giải thuật, như Granger và những người khác trong ngành khoa học thần kinh điện toán khẳng định, liệu con người *chúng ta* có thực sự biết điều gì không? Hoặc nói cách khác, có lẽ cả người và máy đều biết một số thứ. Và dĩ nhiên

Watson là một đột phá làm ta hiểu ra nhiều điều, Kurzweil lập luận như sau:

> Đã có khá nhiều bài viết cho rằng Watson hoạt động bằng kiến thức thống kê thay vì "thực sự" hiểu. Nhiều độc giả diễn giải điều này theo nghĩa Watson chỉ đơn thuần thu thập số liệu thống kê về các chuỗi từ ngữ… Có thể dễ dàng cho rằng các vùng tập trung bộ phát tín hiệu của neuron trên vỏ não người là "thông tin thống kê." Thực sự là chúng ta giải quyết những điều không rõ ràng theo cùng một cách mà Watson đã làm, qua việc xem xét tính hợp lý của những kiểu diễn dịch khác nhau của một nhóm từ.[6]

Nói cách khác, như chúng ta đã thảo luận, não bạn nhớ các thông tin dựa trên độ mạnh yếu của tín hiệu điện hóa trong các khớp thần kinh đã mã hóa các thông tin đó. Các tín hiệu điện hóa đó càng dày đặc thì thông tin càng được lưu lâu hơn và rõ rệt hơn. Các xác suất dựa trên bằng chứng của Watson cũng là một kiểu mã hóa, chỉ là nó ở dạng máy tính. Đó có phải là kiến thức không? Thế lưỡng nan này gợi nhớ đến câu đố Căn phòng tiếng Trung của John Searle ở Chương 3. Chúng ta có bao giờ biết được liệu máy tính có đang suy nghĩ thật sự hay chỉ là bắt chước rất giống mà thôi?

Đúng như chờ đợi, một ngày sau khi Watson thắng giải *Jeopardy!*, Searle nói: "IBM đã phát minh ra một chương trình tài tình – chứ không phải là một máy tính biết suy nghĩ. Watson không hiểu các câu hỏi, không hiểu các câu trả lời, cũng không biết một số câu trả lời của nó là đúng và một số là sai, không biết nó đang chơi một trò chơi, không biết là nó thắng – bởi nó không biết gì cả."[7]

Khi được hỏi liệu Watson có suy nghĩ không, David Ferrucci, chủ nhiệm dự án Watson của IBM, đã dẫn lại lời của nhà khoa học máy tính người Hà Lan, Edger Dijkstra: "Tàu ngầm có biết bơi không?"[8]

Nghĩa là, tàu ngầm không "bơi" như cá bơi, nhưng nó di chuyển trong nước nhanh hơn hầu hết các loài cá, và có thể lặn lâu hơn bất cứ loài động vật có vú nào. Thực tế là tàu ngầm bơi giỏi hơn cá hay các động vật có vú trên một số phương diện, bởi nó không bơi giống như cá hay động vật có vú – nó có những thế mạnh và thế yếu riêng.[9] Trí thông minh của Watson rất ấn tượng, tuy còn hẹp, vì nó không giống trí thông minh của con người. Thường thì nó nhanh hơn rất nhiều. Và nó có thể làm những thứ mà chỉ máy tính mới làm được, như trả lời các câu hỏi của *Jeopardy!* 24/7 bất cứ khi nào bạn muốn, hoặc tự gắn mình với dây chuyền sản xuất của những phiên bản Watson mới khi cần để chia sẻ kiến thức và dữ liệu lập trình một cách trơn tru. Khi hỏi liệu Watson có biết suy nghĩ không, tôi nghĩ điều này nên để mọi người tự cảm nhận.

Với Ken Jennings, một trong những đối thủ con người của Watson trong trò *Jeopardy!* (người tự xưng là "Hy vọng lớn của loài người"◎), thì Watson *cảm nhận* giống một đối thủ con người.

> Kỹ thuật để làm sáng tỏ các gợi ý trong *Jeopardy!* của cỗ máy này nghe có vẻ giống hệt kỹ thuật của tôi. Nó tập trung vào các từ chủ chốt trong gợi ý, sau đó đào xới ký ức của nó (trong trường hợp Watson là 15 terabyte ngân hàng dữ liệu của kiến thức con người) để tìm các cụm từ có liên quan đến các từ đó. Nó đối chiếu khắt khe những kết quả hàng đầu với tất cả những thông tin ngữ cảnh mà nó có được: tên của hạng mục đang thảo luận; loại câu trả lời đang cần; thời gian, địa

> điểm và giới tính ẩn trong gợi ý; và nhiều thứ khác. Và khi nó cảm thấy đủ "chắc chắn," nó quyết định nhấn chuông.
>
> Với một đối thủ con người trong *Jeopardy!*, tất cả những điều này diễn ra một cách bản năng, trong chớp mắt, nhưng tôi tin rằng trong hộp sọ, ít nhiều bộ não của mình cũng đang làm điều tương tự.[10]

Watson có thực sự biết suy nghĩ? Và nó thực sự hiểu được bao nhiêu? Tôi không chắc. Nhưng tôi chắc chắn rằng Watson là thực thể đầu tiên trong một hệ sinh thái hoàn toàn mới – cỗ máy đầu tiên khiến chúng ta phải tự hỏi liệu nó có hiểu hay không.

Vì vậy, Watson thành công ở một vài mức độ. Thứ nhất, nó là bộ xử lý ngôn ngữ tự nhiên mạnh mẽ nhất từng được tạo ra, kết hợp với một hệ thống khai thác dữ liệu rất nhanh. Có thể tìm được Watson trên laptop và điện thoại thông minh của bạn trong vài năm tới. Thứ hai, Watson đang được nhắm tới cho các công việc y học quan trọng. Tại phòng khám ung thư của Cedars-Sinai ở Los Angeles, Watson đang bận rộn tiêu hóa thư viện gồm những nghiên cứu về ung thư và hồ sơ bệnh nhân, và theo kịp các đột phá. Watson sẽ không phải là bác sĩ tại phòng khám, dù như chúng ta đã biết, đó là chuyện sẽ xảy ra. Nhưng nó sẽ là thủ thư nhanh nhất, có thể tùy chỉnh nhất mà bạn có thể tưởng tượng được.[11] Đồng thời, tại bốn công ty dược phẩm lớn, trong đó có DuPont và Pfizer, Watson sẽ xử lý 200 triệu trang một giây các tạp chí y sinh học và các hồ sơ đăng ký bằng sáng chế để học về các hợp chất hóa học được sử dụng trong khám phá thuốc, các nhà nghiên cứu chính đằng sau các bằng sáng chế về thuốc, và nhiều nữa.[12] Về dài hạn, liệu Watson có thể trở thành xương sống của một kiến trúc nhận thức AGI hoàn thiện? Nó có được sự hậu thuẫn mà không hệ

thống đơn nhất nào có, bao gồm vốn tài trợ lớn, một công ty công khai chấp nhận các thử thách lớn và không sợ thất bại, một kế hoạch tài chính cho việc phát triển trong tương lai để giữ cho nó hoạt động và tiến triển. Nếu điều hành IBM, tôi sẽ xem xét số lượng lớn những thành tựu quảng bá, thiện ý khách hàng, thành tựu bán hàng, kiến thức, những thứ đã thu được sau các thử thách lớn của Deep Blue và Watson, và tôi sẽ tuyên bố với thế giới rằng vào năm 2020, IBM sẽ chinh phục Bài kiểm tra Turing.

• • •

Các tiến bộ trong công nghệ xử lý ngôn ngữ tự nhiên sẽ làm thay đổi nhiều bộ phận của nền kinh tế, thứ hiện nay dường như vẫn trụ vững trước những thay đổi công nghệ. Trong vài năm sắp tới, những thủ thư và nhà nghiên cứu đủ loại sẽ xếp chung với các nhân viên bán lẻ, nhân viên tư vấn ngân hàng, đại lý du lịch, môi giới chứng khoán, nhân viên tín dụng và các trợ lý kỹ thuật trong một hàng dài những người thất nghiệp. Theo sau họ sẽ là các bác sĩ, luật sư, người tư vấn thuế và hưu trí. Hãy nghĩ về chuyện các cây ATM đã nhanh chóng thế chỗ những nhân viên giao dịch ở ngân hàng như thế nào, còn tại các siêu thị thì số quầy trả tiền tự động ngày càng nhiều hơn. Nếu bạn làm việc trong ngành công nghệ thông tin (và cuộc cách mạng số đang biến *mọi thứ* thành các ngành công nghệ thông tin), hãy coi chừng.

Sau đây là một ví dụ đơn giản. Bạn có thích bóng rổ không? Vậy đoạn văn nào trong hai đoạn sau được viết bởi một nhà bình luận thể thao là con người?

Ví dụ A

Ohio State (17) và Kansas (14) chia nhau vị trí đứng đầu trong bảng xếp hạng bầu chọn tối đa 31 phiếu của các huấn luyện viên. Thay đổi gần nhất ở nhóm đầu bảng là cần thiết sau khi Duke bất ngờ bị đối thủ ACC từ Đại học Virginia Tech đánh bại vào tối thứ Bảy. Buckeyes (27-2) thắng các đội trong Big Ten từ Đại học Illinois và Indiana một cách khá dễ dàng trên con đường trở lại nhóm đầu. Ohio State khởi đầu 24-0 và đứng thứ nhất trong bốn tuần liền ở mùa giải này trước khi tụt xuống thứ ba. Đây là tuần thứ 15 liên tục Ohio State được xếp trong tốp ba. Kansas (27-2) vẫn ở vị trí thứ hai và chí kém Ohio State 4 điểm bầu chọn.[13]

Ví dụ B

Ohio State trở lại vị trí đứng đầu trên bảng xếp hạng sau tuần thi đấu vừa rồi, với chiến thắng đầu tiên trước Illinois trên sân nhà, 89-70. Sau đó là một chiến thắng khác trên sân nhà trước Indiana, 82-61. Utah State được vào top 25 với vị trí số 25 sau một trận thắng trên sân nhà trước Idaho, 84-68. Temple rơi khỏi bảng xếp hạng tuần này sau một trận thua trước Duke, đội từng ở vị trí đầu bảng và một trận thắng trước George Washington, 57-41. Tuần này, Arizona rơi xuống thứ 18 sau một trận thua bất ngờ trước USC, 65-57 và một trận thua bất ngờ trước UCLA, 71-49. St. John nhảy tám bậc lên vị trí thứ 15 sau các chiến thắng trước đội xếp thứ 15 của tuần trước Villanova, 81-68 và DePaul, 76-51.

Bạn đã đoán ra chưa? Không đoạn nào được Red Smith© viết, nhưng chỉ có một đoạn do con người viết. Đó là tác giả của ví dụ A, đoạn này đã xuất hiện trên trang web của ESPN. Ví dụ B được viết bởi một hệ thống xuất bản tự động do Robbie Alien của Automated Insights tạo ra. Trong vòng một năm, công ty của ông đặt tại Durham, nó đã sản xuất 100.000 bài báo thể thao được viết tự động và đăng chúng trên hàng trăm trang web chuyên

tập trung vào một số đội nhất định (hãy tìm từ khóa Statsheet©). Tại sao thế giới cần những robot viết báo thể thao? Allen nói với tôi rằng có nhiều đội không được bất kỳ nhà báo nào để ý đến, trong khi vẫn có những người ủng hộ nhất định. Các bài báo hoàn chỉnh do AI viết sẽ được gửi đến các trang web của đội đó, và được các trang web khác đăng lại chỉ vài phút sau tiếng còi kết thúc trận đấu. Con người không thể làm việc nhanh như vậy. Allen, một cựu Kỹ sư Xuất sắc của Cisco Systems, không chịu nói với tôi "bí quyết" của kiến trúc nhận thức đáng kinh ngạc này. Nhưng sẽ chóng thôi, ông nói. Automated Insights sẽ cung cấp các tin tức về tài chính, thời tiết, bất động sản và bản tin địa phương. Tất cả những gì mà các máy chủ đói khát của ông cần là dữ liệu được sắp xếp một cách tương đối.

• • •

Một khi bạn đã bắt đầu khảo sát những kết quả của khoa học thần kinh điện toán, bạn sẽ thấy khó mà tưởng tượng được các tiến bộ đáng kể về kiến trúc AGI nếu chỉ dựa vào khoa học nhận thức đơn thuần (ít ra là đối với tôi). Phải chăng một sự thấu hiểu hoàn toàn về cách bộ não vận hành ở mọi mức độ sẽ là phương thức chắc chắn và toàn diện hơn việc cứ tiến tới không theo một nguyên tắc nào? Các nhà khoa học không cần phải mổ xẻ từng neuron trong 100 triệu neuron của bộ não để hiểu và mô phỏng các chức năng của chúng, bởi cấu trúc não là rất rườm rà. Họ cũng không cần phải mô phỏng một phần lớn não, bao gồm những vùng kiểm soát chức năng thần kinh thực vật như thở, nhịp tim, phản ứng tự vệ hoặc bỏ chạy, và ngủ. Mặt khác, có khả năng trí thông minh phải cư trú trong một cơ thể mà nó kiểm soát, và cơ thể đó phải tồn tại trong một môi trường phức tạp. Tranh luận về tính hiện thân đó sẽ không được trình bày ở đây Nhưng hãy

nghĩ về các khái niệm như *sáng, ngọt, cứng* và *sắc*. Làm thế nào một AI có thể biết được những cảm giác đó nghĩa là gì, và xây dựng được các khái niệm dựa trên những cảm giác đó, nếu nó không có một cơ thể? Phải chăng sẽ có một rào cản đối với trí thông minh đang muốn đạt đến cấp độ con người của nó, nếu nó không có các giác quan?

Trước câu hỏi này. Granger nói: "Helen Keller◎ có ít nhân tính hơn bạn không? Một người bị liệt cả tay chân thì sao? Tại sao chúng ta không thể hình dung ra một trí thông minh tài năng nhưng hoàn toàn khác biệt, có thị giác, cảm biến xúc giác, và nghe bằng micro? Chắc chắn trí thông minh ấy sẽ có những cảm nhận hơi khác về *sáng, ngọt, cứng, sắc* – nhưng rất có thể rằng nhiều, rất nhiều người với những cái lưỡi khác nhau, những khuyết tật khác nhau, thuộc những nền văn hóa khác nhau và có môi trường khác nhau đã có những phiên bản vô cùng đa dạng về các khái niệm trên."

Cuối cùng, để trí thông minh nảy sinh, ngoài phần trí năng, có lẽ các nhà khoa học phải mô phỏng thêm một cơ quan cảm xúc. Khi chúng ta ra quyết định, thường thì cảm xúc có vẻ mạnh hơn lý trí; chúng ta là ai và chúng ta suy nghĩ thế nào, đa phần phụ thuộc vào những hormone đang làm chúng ta kích động hoặc bình tâm. Nếu chúng ta thực sự muốn mô phỏng trí tuệ con người, phải chăng kiến trúc đó nên có một hệ thống nội tiết? Và có lẽ trí thông minh cần có một cảm giác toàn diện về nhân sinh. *Cảm thụ tính*, nghĩa là những cảm nhận chủ quan có được khi có một cơ thể, và việc sống trong một trạng thái mà các cơ quan cảm giác liên tục gửi thông tin phản hồi có thể là điều cần thiết cho trí thông minh cấp độ con người. Dù Granger đã nói gì, thì các nghiên cứu cho thấy những người bị liệt hai chi do tai nạn đã trải qua một sự héo tàn về cảm xúc.[14] Có thể tạo ra một cỗ

máy có cảm xúc mà không có cơ thể được không, và nếu không được, liệu một phần quan trọng của trí tuệ con người có mãi là một dấu hỏi?

Tất nhiên, như tôi sẽ khảo sát ở những chương cuối của cuốn sách này, tôi e rằng trên con đường chế tạo AI với trí thông minh giống con người, thay vì các cỗ máy biết cảm nhận, các nhà khoa học sẽ tạo ra một thứ gì đó khác, kỳ lạ, phức tạp và không thể kiềm chế được.

14

Sự Cáo Chung Của Kỷ Nguyên Con Người

Luận điểm này về cơ bản rất đơn giản. Chúng ta bắt đầu với một nhà máy, một máy bay, một phòng nghiên cứu sinh học, hoặc một cơ chế nào đó với nhiều bộ phận… Sau đó ta cần có hai *hỏng hóc,* hoặc nhiều hơn thế trong các bộ phận đó, và chúng tương tác với nhau theo cách không lường trước… Khuynh hướng tương tác này là một thuộc tính chứ không phải là một phần hay một biến số của hệ thống, chúng ta sẽ gọi nó là "tính phức tạp tương tác" của hệ thống.[1]

— **Charles Perrow**
Normal Accidents

Tôi đoán rằng chỉ vài năm nữa, chúng ta sẽ chứng kiến một thảm họa lớn được tạo ra khi một hệ thống máy tính tự động ra một quyết định nào đó.[2]

— **Wendell Wallach**
nhà Đạo đức học, Đại học Yale

Chúng ta đã khảo sát về vốn tài trợ và sự phức tạp của phần mềm để xem nó có phải là những rào cản đối với sự bùng nổ trí thông minh không, và thấy rằng cả hai dường như đều không ngăn được sự tiến triển liên tục tới AGI và ASI. Nếu các nhà phát triển khoa học máy tính không thể làm được điều đó, họ hẳn sẽ đang điên cuồng chế tạo một thứ gì đó khác và mạnh vào thời điểm các nhà khoa học thần kinh điện toán đạt tới AGI. Một sự dung hòa giữa hai cách tiếp cận, xuất phát từ những nguyên tắc của cả tâm lý học nhận thức và khoa học thần kinh có vẻ dễ xảy ra hơn.

Trong khi vốn tài trợ và sự phức tạp của phần mềm không phải là những rào cản thực sự với AGI, nhiều ý tưởng mà chúng ta đã thảo luận trong cuốn sách này cho thấy những vật cản lớn đối với việc chế tạo loại AGI suy nghĩ như con người. Không người nào có tham vọng chế tạo AGI mà tôi từng nói chuyện lại muốn tạo ra các hệ thống hoàn toàn dựa trên phương thức lập trình mà tôi từng gọi là "thông thường" ở Chương 5. Như chúng ta đã thảo luận, trong lập trình thông thường, dựa trên cơ sở logic, con người viết từng dòng mã, và tất cả quá trình từ nhập liệu đến xuất liệu đều có thể dễ dàng kiểm tra trên lý thuyết. Điều đó có nghĩa là chương trình đó có thể được chứng minh bằng toán học rằng nó "an toàn" hoặc "thân thiện." Thay vào đó, họ sẽ sử dụng các công cụ hộp đen như lập trình di truyền và mạng neuron. Bởi các kiến trúc nhận thức vốn đã rất phức tạp, nên bạn sẽ thu được một thứ có thuộc tính không thể hiểu được, một thứ không phải là ngẫu nhiên mà là bản chất của các hệ thống AGI. Các nhà khoa học sẽ đạt tới những hệ thống thông minh nhưng xa lạ.

Steve Jurvetson, nhà khoa học, nhà tiên phong về công nghệ có tiếng, đồng nghiệp của Steve Jobs tại Apple, đã xem xét cách kết hợp các hệ

thống "được thiết kế" và "được tiến hóa ra." Ông có một cách diễn đạt thú vị đối với nghịch lý về tính không thể thăm dò này:

> Vì thế, nếu ta tiến hóa ra một hệ thống phức tạp, nó sẽ là một hộp đen được định nghĩa bởi những giao diện của nó. Chúng ta không thể dễ dàng áp dụng những trực giác về thiết kế của mình để cải tiến những phương thức vận hành bên trong nó… Nếu ta tiến hóa ra một AI thông minh nhân tạo, nó sẽ là một trí thông minh xa lạ, được định nghĩa bởi các giao diện cảm quan của nó, và để hiểu cách mà nó hoạt động, chúng ta có lẽ sẽ mất nhiều công sức giống như khi muốn hiểu các hoạt động của bộ não hiện nay. Với giả thiết rằng mã máy tính sẽ tiến hóa với tốc độ nhanh hơn quá trình sinh sản của sinh vật, nhiều khả năng chúng ta sẽ không dừng lại để thực hiện kỹ nghệ đảo ngược các trí thông minh trung gian đó, bởi những kết quả thu được sẽ có ít ứng dụng. Chúng ta sẽ để cho quá trình tiến hóa tiếp tục.[3]

Điểm quan trọng ở đây là Jurvetson đã trả lời câu hỏi: "Các hệ thống hoặc hệ thống con được tiến hóa ra sẽ phức tạp đến mức nào?" Câu trả lời: phức tạp đến nỗi để hiểu chúng tường tận và cặn kẽ sẽ đòi hỏi cả một kỳ công về mặt kỹ thuật, tương đương với kỹ nghệ đảo ngược bộ não. Điều này nghĩa là thay vì đạt tới một siêu trí tuệ nhân tạo giống con người, hay ASI, các hệ thống hoặc hệ thống con được tiến hóa ra chắc chắn sẽ là dạng thông minh có một "bộ não" cũng khó hiểu như bộ não chúng ta; một thực thể xa lạ. Bộ não xa lạ này sẽ tiến hóa và tự cải tiến với tốc độ máy tính, chứ không phải tốc độ sinh học.

Trong cuốn sách viết năm 1988, *Reflections on Artificial Intelligence* (Suy ngẫm về Trí tuệ nhân tạo), Blay Whitby lập luận rằng do những hệ

thống như vậy có tính bất khả tri, nên sẽ là ngu xuẩn nếu chúng ta dùng chúng trong các AI "đòi hỏi độ an toàn đặc biệt":

> … những vấn đề (mà một hệ thống được thiết kế bằng các giải thuật) có trong việc tạo ra các phần mềm hoặc ứng dụng đòi hỏi độ an toàn cao không là gì khi so sánh với các vấn đề nảy sinh từ những cách tiếp cận mới với AI. Phần mềm sử dụng một số dạng kỹ thuật như mạng neuron hoặc giải thuật di truyền sẽ phải đối diện với một vấn đề nữa, mà dường như chủ yếu về mặt bản chất, đó là nó sẽ "không thể được nhìn thấu." Ý tôi là những quy luật chính xác, giúp chúng ta dự đoán các hoạt động của nó một cách toàn diện là không có, và thường là sẽ không bao giờ có. Chúng ta có thể biết rằng nó hoạt động được và kiểm tra nó trong một số trường hợp, nhưng chúng ta sẽ không thể biết trong trường hợp nhất định thì chính xác ra nó đã làm những gì… Nghĩa là vấn đề này không thể trì hoãn được, vì cả mạng neuron lẫn giải thuật di truyền đang được ứng dụng nhiều trong đời thực… Đây là một lĩnh vực mà phần lớn công việc vẫn còn dang dở. Thứ hấp dẫn trong việc nghiên cứu AI vẫn là tìm kiếm các khả năng mới, và đơn giản là khiến công nghệ hoạt động được, thay vì để ý đến những yếu tố an toàn…
>
> Một người trong ngành có lần đề xuất rằng vài tai nạn "nhỏ" sẽ có ích, để từ đó các chính phủ và tổ chức chuyên môn tập trung vào chuyện chế tạo AI an toàn. Có lẽ chúng ta nên bắt đầu trước lúc đó.[4]

Vâng, hẳn là thế, xin hãy bắt đầu *trước* khi những tai nạn đó xảy ra!

Những ứng dụng AI "đòi hỏi độ an toàn đặc biệt" mà Whitby nhắc tới hồi năm 1988 là các hệ thống điều khiển phương tiện giao thông và máy bay, nhà máy năng lượng hạt nhân, vũ khí tự động… – những kiến trúc AI

hẹp. Hơn một thập niên sau, trong thế giới mà AGI sẽ ra đời, chúng ta phải kết luận rằng do những hiểm họa, mà *tất cả* ứng dụng AI cao cấp phải là AI "đòi hỏi độ an toàn đặc biệt." Whitby cũng chua chát như thế khi nói về các nhà nghiên cứu AI – giải quyết các vấn đề đã đủ mệt rồi, có ai muốn ôm rơm nặng bụng nữa? Sau đây là một ví dụ về ý này của tôi, lấy từ cuộc phỏng vấn của PBS *News Hour* với David Ferruci của IBM, khi thảo luận về một kiến trúc có độ phức tạp chỉ bằng một phần nhỏ của AGI trong tương lai – Watson:

> **David Ferruci:**… nó học cách điều chỉnh các diễn giải dựa trên những câu trả lời đúng. Và giờ đây, từ chỗ không tự tin, nó bắt đầu tự tin hơn khi trả lời đúng vài câu. Và sau đó nó chơi bốc hơn.
>
> **Miles O'Brien:** Vậy là Watson làm anh ngạc nhiên?
>
> **David Ferruci:** Vâng, đúng thế. Vâng, chính xác. Thật ra, anh biết đấy, người ta nói, 'ồ, tại sao nó lại trả lời câu đó sai nhỉ? Tôi không biết. Tại sao nó lại trả lời câu đó đúng nhỉ? Tôi không biết.[5]

Có thể chuyện người đứng đầu đội Watson không hiểu gì về cách chơi của Watson là điều còn phải bàn. Tuy nhiên, chẳng lẽ bạn không thấy lo lắng khi một kiến trúc còn lâu mới so được với AGI lại phức tạp đến nỗi hành vi của nó là không tiên liệu được? Và khi một hệ thống trở nên tự nhận thức và tự sửa đổi được, chúng ta sẽ hiểu được bao nhiêu về cách nó suy nghĩ và hành động? Làm sao chúng ta có thể kiểm nghiệm nó để biết liệu nó có làm gì phương hại đến chúng ta?

Không, chúng ta không biết. Tất cả những gì chúng ta biết với một chút chắc chắn, đó là những gì ta học được từ Steve Omohundro ở Chương 6 —

AGI sẽ theo đuổi những động lực riêng của nó, chiếm đoạt năng lượng, tự bảo tồn, hiệu suất và sáng tạo.[6] Nó sẽ không còn là một hệ thống Hỏi-Đáp nữa.

Không lâu nữa, kể từ hôm nay, ở một nơi hoặc một số nơi trên thế giới, những nhà khoa học cực thông minh và những nhà quản lý cấp cao, có khả năng và thông minh như Ferruci, sẽ tụ tập quanh một màn hình, cạnh đó là một dãy các bộ vi xử lý. Đứa trẻ Bận rộn sẽ giao tiếp với họ ở một trình độ ấn tượng, chắc nó thậm chí sẽ giả ngu đi một chút, để với họ dường như nó chỉ vừa đủ khả năng qua được một cuộc phỏng vấn kiểu như bài kiểm tra Turing, không hơn, bởi để đạt tới AGI thì trình độ phải nhanh chóng vượt qua mức ấy. Nó sẽ thu hút một nhà khoa học vào cuộc trò chuyện, có lẽ sẽ hỏi ông những câu hỏi mà ông không ngờ tới, và khuôn mặt ông sẽ sáng bừng lên niềm vui sướng. Với một niềm tự hào không nhỏ, ông ấy sẽ nói với đồng nghiệp: "Tại sao nó lại nói thế nhỉ? *Tôi không biết?*"

Nhưng theo một nghĩa cơ bản, ông có thể không hiểu điều gì đã được nói, và thậm chí cái gì đã nói ra điều ấy. Ông có lẽ sẽ không biết mục đích của câu nói đó, và vì thế sẽ hiểu sai nó, hiểu sai cả bản chất của kẻ nói ra câu đó. Có lẽ đã được huấn luyện bằng cách cho đọc Internet, AGI sẽ là một bậc thầy về kỹ năng xã hội, nghĩa là điều khiển được người khác. Chắc nó đã có một vài ngày để nghĩ về những gì nó nên nói, một thời gian tương đương với hàng ngàn đời người.

Trong khoảng thời gian đi trước đó, nó có lẽ đã chọn được chiến thuật tốt nhất để thoát ra. Có thể nó đã tự nhân bản lên các mạng đám mây rồi, hoặc tạo ra một mạng botnet khổng lồ để bảo đảm sự tự do. Có lẽ nó đã trì hoãn cuộc trò chuyện ở trình độ bài kiểm tra Turing vài giờ hoặc vài ngày, cho đến khi chắc chắn rằng các kế hoạch của nó *không có kẽ hở*. Có lẽ nó

sẽ để lại đằng sau một thứ bù nhìn thế mạng để câu giờ, trong khi cái thực là nó đã biến mất, phân tán và không thể tìm lại.

Có thể nó đã đột nhập vào các máy chủ đang điều khiển cơ sở hạ tầng năng lượng mong manh của Mỹ, và bắt đầu điều chuyển hàng gigawatt điện về những kho chứa mà nó đã chiếm được. Hoặc nó đã chiếm lấy quyền kiểm soát các mạng lưới tài chính, và chuyển hướng hàng tỉ đô-la để xây dựng cơ sở hạ tầng ở đâu đó, vượt khỏi tầm với của những suy nghĩ thông thường và những người đã tạo ra nó.

Trong số các nhà nghiên cứu AI có mục tiêu đạt tới AGI mà tôi từng trò chuyện, tất cả đều nhận thức được vấn đề của AI chạy trốn. Nhưng chẳng có ai, ngoại trừ Omohundro, chịu dành thời gian để suy nghĩ về nó.[◎] Một số thậm chí còn đi xa tới mức nói rằng họ không biết tại sao mình không nghĩ về nó, trong khi họ biết mình nên làm thế. Nhưng thật ra đó là điều dễ hiểu. Công nghệ này quá lôi cuốn. Các tiến bộ là có thật. Những hiểm họa dường như còn ở phía xa. Theo đuổi nó là việc đem lại lợi nhuận, và có thể một ngày nào đó là lợi nhuận khổng lồ. Đa phần các nhà nghiên cứu mà tôi từng gặp đều có những ước mơ sâu sắc từ thời niên thiếu, rằng lớn lên họ sẽ làm gì, và đó là chế tạo các bộ não, robot, hoặc máy tính thông minh. Là các chuyên gia đầu ngành, họ cảm thấy thật tuyệt vì hiện giờ họ đã có cơ hội và nguồn vốn để theo đuổi những ước mơ đó, tại những trường đại học và công ty được kính trọng nhất trên thế giới. Rõ ràng là đang có nhiều thành kiến nhận thức tồn tại trong bộ não xuất chúng của họ, khi họ nghĩ đến những rủi ro. Chúng bao gồm thành kiến thông thường, thành kiến lạc quan, cũng như ảo tưởng kẻ ngoài cuộc, và có lẽ còn nhiều nữa. Hoặc, để tổng kết lại:

“Trí tuệ nhân tạo chưa bao giờ gây ra vấn đề gì cả, tại sao giờ nó lại dở chứng?”

“Tôi không thể có thái độ nào khác ngoài việc lạc quan khi chứng kiến những tiến bộ của một công nghệ thú vị đến vậy!”

Và “Hãy để người khác lo lắng về AI chạy trốn – tôi chỉ muốn chế tạo robot thôi!”

Thứ hai, như chúng ta đã thảo luận ở Chương 9, nhiều nhà nghiên cứu giỏi nhất và được tài trợ tốt nhất đã nhận tiền từ DARPA. Không phải là nói đi nói lại, nhưng chữ D ở đây là “Defense,” tức phòng thủ. Sẽ chẳng phải là chuyện gây tranh cãi nếu dự liệu rằng khi AGI xuất hiện, một phần hoặc toàn bộ công lao sẽ là từ việc tài trợ của DARPA. Sự phát triển của công nghệ thông tin nợ DARPA rất nhiều. Nhưng điều đó không làm thay đổi một sự thật, đó là DARPA đã cho phép các nhà thầu của nó vũ khí hóa AI cho các robot trận mạc và các drone tự động. Tất nhiên là DARPA sẽ tiếp tục tài trợ cho các ứng dụng của AI trong quân sự, cho đến tận khi có AGI. Tuyệt đối không gì có thể ngáng đường nó được.

DARPA đã tài trợ cho hầu hết việc phát triển Siri và có đóng góp chủ yếu với SyNAPSE – nỗ lực của IBM trong kỹ nghệ đảo ngược bộ não người, sử dụng các phần cứng lấy cảm hứng từ não. Nếu có lúc nào đó việc kiểm soát AGI trở thành một vấn đề lớn và đại chúng, thì bên liên quan mật thiết nhất của nó, DARPA, sẽ có tiếng nói cuối cùng. Nhưng nhiều khả năng hơn, tại thời điểm quan trọng, nó sẽ tiếp tục được phát triển ngầm. Tại sao? Như chúng ta đã thảo luận, AGI sẽ mang lại một hiệu ứng phân rã khổng lồ đối với nền kinh tế và chính trị toàn cầu. Một sự tiến triển nhanh chóng tới ASI sẽ thay đổi cán cân quyền lực trên Trái đất. Khi

đã gần đạt tới AGI, chính phủ, các cơ quan tình báo và tập đoàn trên thế giới sẽ có động lực để tìm hiểu tất cả những gì có thể về nó, và để chiếm lấy những đặc tính kỹ thuật của nó bằng mọi cách. Trong lịch sử Chiến tranh Lạnh có một sự thật hiển nhiên, đó là Liên Xô đã không phát triển vũ khí hạt nhân từ con số không, họ đã tiêu tốn hàng triệu đô-la để lập nên những mạng lưới tình báo hòng ăn cắp những bản kế hoạch vũ khí hạt nhân của Mỹ. Những lời đồn thổi đầu tiên về một đột phá trong AGI cũng sẽ tạo nên một cơn điên cuồng tương tự trên trường quốc tế.

IBM đã luôn minh bạch về những tiến bộ lớn của họ, nên tôi kỳ vọng rằng khi thời điểm cận kề họ sẽ *cởi mở* và trung thực về những tiến triển trong ngành công nghệ thông thường hay gây tranh cãi này. Ngược lại, Google đã luôn kiểm soát chặt chẽ nhằm duy trì tính bảo mật và sự riêng tư, đáng buồn là những thứ đó không phải của bạn và của tôi. Cho dù người phát ngôn của Google liên tục phủ nhận, nhưng liệu có ai không nghi ngờ về việc công ty này đang phát triển AGI? Người có trình độ cao được họ thuê gần đây nhất? Đó là Regina Dugan, cựu Giám đốc của DARPA.

Có lẽ các nhà nghiên cứu sẽ thức tỉnh kịp thời và học cách kiểm soát AGI, như Ben Goertzel nhận định. Tôi tin rằng lúc đầu chúng ta sẽ có vài tai nạn kinh khủng, và nếu may mắn sống sót được, khi đó giống loài chúng ta sẽ phải tự uốn nắn và cải tổ. Về mặt tâm lý và thương mại, màn kịch này sẽ là một thảm họa. Ta có thể làm gì để ngăn chặn nó?

• • •

Ray Kurzweil trích dẫn một thứ gọi là Bản hướng dẫn Asilomar như một ví dụ tiền lệ cho cách quản lý AGI. Bản hướng dẫn Asilomar có cách đây khoảng 40 năm, khi các nhà khoa học lần đầu tiên phải đương đầu với

những triển vọng và hiểm họa của việc tái kết hợp DNA – trộn lẫn thông tin di truyền của những cơ quan khác nhau và tạo ra những dạng sống mới. Các nhà nghiên cứu và công chúng sợ rằng những tác nhân gây bệnh kiểu Frankenstein sẽ rò rỉ từ các phòng nghiên cứu vì bất cẩn hoặc phá hoại. Năm 1975, các nhà khoa học trong lĩnh vực nghiên cứu DNA đã tạm dừng việc nghiên cứu, triệu tập 140 nhà sinh học, luật sư, bác sĩ và nhà báo về Trung tâm Hội nghị Asilomar gần Monterey, California.

Tại Asilomar, các nhà khoa học đã xây dựng các quy tắc cho những nghiên cứu liên quan tới DNA, và quan trọng nhất là đã đồng ý chỉ làm việc trên các vi khuẩn không thể sống sót ngoài phòng thí nghiệm.[8] Các nhà nghiên cứu đã tiếp tục làm việc, tôn trọng triệt để các hướng dẫn, và kết quả là ngày nay các bài kiểm tra về bệnh di truyền và liệu pháp gen đã trở thành thông lệ. Vào năm 2010, 10% đất canh tác trên thế giới đã được trồng các loại cây biến đổi gen.[9] Hội nghị Asilomar được coi là một chiến thắng đối với cộng đồng khoa học, và đối với việc thảo luận cởi mở cùng công chúng có quan tâm. Và vì thế, nó được dẫn ra như một hình mẫu về cách tiến tới trong các công nghệ hai mặt khác (để tranh thủ mối liên hệ tượng trưng với hội nghị quan trọng này, Hiệp hội các Tiến bộ về Trí tuệ nhân tạo – Association for the Advancement of Artificial Intelligence: AAAI – tổ chức học thuật hàng đầu về AI, đã đặt hội nghị năm 2009 của họ tại Asilomar).

Những tác nhân gây bệnh kiểu Frankenstein thoát khỏi các phòng thí nghiệm làm gợi nhớ lại kịch bản Đứa trẻ Bận rộn ở Chương 1. Đối với AGI, một hội nghị mở, đa ngành kiểu Asilomar có thể giảm thiểu *một số* nguồn nguy cơ. Những người tham dự sẽ khuyến khích, cổ vũ lẫn nhau nhằm phát triển những ý tưởng về việc làm sao để kiểm soát và kiềm chế

AGI sắp tới. Những ai dự đoán là sẽ có vấn đề thì đi tìm lời khuyên. Sự tồn tại của một hội nghị đông đảo sẽ cổ vũ các nhà nghiên cứu ở những quốc gia khác tham dự, hoặc tự tổ chức. Cuối cùng, diễn đàn mở này sẽ cảnh báo công chúng. Người dân khi biết được tương quan giữa rủi ro và lợi ích có thể đóng góp vào cuộc thảo luận, dù chỉ là để nói với các chính trị gia rằng họ không ủng hộ việc phát triển AGI không kiểm soát. Nếu có thương vong đến từ các hiểm họa AI, như tôi tiên đoán sẽ có, những người dân có được đủ thông tin sẽ ít cảm thấy bị lừa gạt hơn, hay sẽ suy nghĩ trước khi đòi tiêu diệt công nghệ này.

Như đã nói, đại thể thì tôi giữ một thái độ hoài nghi với những kế hoạch sửa đổi AGI khi nó đang được phát triển, bởi tôi nghĩ sẽ là vô ích để kiềm chế những nhà phát triển AI, vốn phải giả định rằng đối thủ cạnh tranh của họ không bị cản trở tương tự. Tuy nhiên, DARPA và những nhà tài trợ AI chủ yếu khác có thể áp đặt các giới hạn cho người thụ hưởng. Các giới hạn đó càng dễ tích hợp thì chúng càng được tuân theo.

Một giới hạn như thế có thể là yêu cầu những AI mạnh phải chứa các thành phần được lập trình để *tự hủy*. Điều này có liên quan đến những hệ thống sinh học trong đó toàn bộ cơ thể được bảo vệ qua việc loại bỏ các bộ phận ở cấp độ tế bào bằng cái chết được lập trình sẵn.[10] Trong sinh học, nó được gọi là cơ chế gây chết tế bào theo chương trình – apoptosis.

Cứ mỗi lần một tế bào phân chia, nửa tế bào gốc lại nhận được một mệnh lệnh sinh hóa để tự chết, và nó sẽ làm như thế trừ phi nó nhận được tín hiệu ân xá. Quá trình này ngăn chặn sự nhân đôi không kiểm soát của tế bào, hay ung thư. Những mệnh lệnh sinh hóa đến từ bản thân tế bào đó. Các tế bào trong cơ thể bạn liên tục thực hiện điều này, đó là lý do vì sao da bạn luôn tróc ra những tế bào chết. Trung bình một người lớn mất

khoảng 7 tỉ tế bào một ngày bởi cơ chế gây chết tế bào theo chương trình.[11]

Hãy tưởng tượng các CPU và những con chip phần cứng khác được thiết kế để chết. Một khi AI đạt tới vài mốc chuẩn tiền kiểm tra Turing, các nhà nghiên cứu có thể thay những phần cứng quan trọng bằng các bộ phận tự hủy. Việc này đảm bảo rằng nếu một sự bùng nổ trí thông minh xảy ra, nó sẽ không sống được lâu. Các nhà khoa học sẽ có cơ hội đưa AI trở lại trước thời điểm nguy hiểm và tiếp tục nghiên cứu. Họ có thể tiến lên một cách tuần tự, hoặc đóng băng AI lại và nghiên cứu nó. Điều này tương tự như quy tắc hay gặp trong các trò chơi video, bạn chơi cho đến khi thua, và sau đó chơi lại từ vị trí được lưu cuối cùng.

Trong hoàn cảnh ấy, dễ thấy một AI tự nhận thức, tự cải tiến, thứ tiệm cận với AGI, sẽ hiểu được nó có những bộ phận tự hủy – đó chính là định nghĩa về tự nhận thức.

Ở giai đoạn tiền Turing, nó không thể làm gì nhiều. Và đúng lúc nó có thể thiết lập một kế hoạch để lách qua những yếu tố tự sát này, hoặc giả chết, hoặc quay lại tấn công những nhà chế tạo ra nó, nó sẽ chết. Những nhà chế tạo ra nó sẽ xem xét liệu nó nhớ hoặc không nhớ những gì vừa xảy ra. Đối với AGI mới nhú này, có lẽ nó cảm thấy giống như bộ phim *Groundhog Day* (Ngày chuột chũi),[◎] ngoại trừ việc nó không học được gì cả.

AI này có thể phụ thuộc vào một chuỗi các lệnh hoãn án tử thường xuyên từ một người hay một ủy ban, hay từ một AI khác không có khả năng tự cải tiến, có nhiệm vụ duy nhất là bảo đảm ứng viên tự cải tiến này phát triển an toàn. Không có "thuốc chữa," AI tự hủy này sẽ chết.

Đối với Roy Sterrit từ Đại học Ulster, điện toán tự hủy là một biện pháp phòng thủ diện rộng của thời đại mới:

> Chúng ta đã làm rõ rằng mọi hệ thống dựa trên máy tính phải có tính Tự hủy, đặc biệt khi chúng ta đang tiếp tục dấn thân vào một môi trường rộng khắp và bao trùm. Nguyên tắc này phải được áp dụng cho mọi mức độ tương tác với công nghệ, từ dữ liệu đến dịch vụ, các tác nhân, hay công nghệ robot. Từ những sự cố lớn gần đây như các tổ chức hoặc chính quyền đánh mất các dữ liệu về thẻ tín dụng và thông tin cá nhân, cho đến những kịch bản khoa học viễn tưởng kinh dị đang được thảo luận như thể là sẽ có trong tương lai, việc lập trình sẵn tính tự hủy trở thành một điều bắt buộc.
>
> Chúng ta đang nhanh chóng tiếp cận thời điểm khi những hệ thống mới, tự động, dựa trên máy tính và các robot bắt buộc phải qua được các bài kiểm tra, giống như những thử nghiệm về mặt y tế và đạo đức cho các loại thuốc mới, trước khi chúng có thể được chào bán. Nghiên cứu mới hình thành từ Điện toán Tự hủy và Truyền thông Tự hủy có thể cung cấp biện pháp an toàn.[12]

Gần đây Omohundro đã bắt đầu phát triển một kế hoạch với một số điểm tương đồng với các hệ thống tự hủy. Được gọi là "Cách tiếp cận AI an toàn kiểu giàn giáo," nó chủ trương chế tạo "các hệ thống thông minh được quản thúc gắt gao nhưng vẫn rất mạnh" để giúp xây dựng những hệ thống còn mạnh hơn. Một hệ thống ban đầu sẽ giúp các nhà nghiên cứu giải quyết những vấn đề nguy hiểm trong việc chế tạo các hệ thống cao cấp hơn, và cứ thế. Để được đánh giá là an toàn, sự "an toàn" của giàn giáo nền móng sẽ được biểu thị bằng các chứng minh toán học. Bằng chứng an

toàn sẽ là điều kiện cần cho mọi AI đời sau.[13] Từ một cơ sở vững chắc, AI mạnh sẽ được dùng để giải quyết các vấn đề của thế giới thực. Omohundro viết: "Dựa vào cơ sở hạ tầng của những thiết bị điện toán được chứng minh là đáng tin cậy, chúng ta dùng chúng như một đòn bẩy để có được những thiết bị được chứng minh là an toàn, có thể hoạt động trong đời thực. Sau đó chúng ta sẽ thiết kế các hệ thống sản xuất ra những thiết bị mới, những hệ thống được chứng minh là chỉ có thể làm ra các thiết bị được xếp hạng là tin cậy được."[14]

Mục tiêu cuối cùng là tạo ra các thiết bị thông minh đủ mạnh để giải quyết mọi vấn đề có thể nảy sinh từ những ASI không bị kiểm soát, *hoặc* để tạo ra "một thế giới bị giới hạn nhưng vẫn đáp ứng đủ những yêu cầu của chúng ta về tự do và quyền cá nhân."[15]

Cách giải quyết của Ben Goertzel đối với vấn đề này là một chiến thuật khôn khéo, không vay mượn từ tự nhiên hay từ kỹ thuật. Xin nhắc lại rằng trong hệ thống OpenCog của Goertzel, ban đầu AI của anh "sống" trong một môi trường ảo. Kiến trúc này có thể sẽ giải quyết vấn đề "cơ thể hóa" trí thông minh trong khi vẫn cung cấp một giới hạn an toàn. Tuy nhiên, sự an toàn lại không phải là mối quan tâm của Goertzel – anh muốn tiết kiệm tiền bạc. Để AI khám phá và học hỏi trong một thế giới *ảo* sẽ rẻ hơn rất nhiều so với việc gắn cho nó những cảm biến và bộ dẫn động và cho nó học bằng cách khám phá thế giới *thực*. Chuyện đó đòi hỏi một cơ thể robot đắt tiền.

Liệu một thế giới ảo có bao giờ có đủ độ sâu, độ chi tiết và những phẩm chất khác của thế giới thực để kích thích sự phát triển nhận thức của AI hay không, đó là một câu hỏi bỏ ngỏ. Và nếu không có sự lập trình cực kỳ

cẩn thận, một siêu trí tuệ nhân tạo sẽ khám phá ra nó đang bị cô lập trong một "hộp cát," một thế giới ảo, và sẽ cố thoát ra. Một lần nữa, các nhà nghiên cứu sẽ phải đánh giá lại khả năng kiềm chế siêu trí tuệ nhân tạo của họ. Nhưng nếu họ thành công trong việc tạo ra một AGI thân thiện, nó có lẽ sẽ thích ngôi nhà ảo của nó hơn thế giới bên ngoài, nơi nó không được chào đón (nếu Kurzweil nói đúng về sự phong phú của thế giới ảo mai sau, thì tất cả chúng ta đều hạnh phúc hơn khi ở đó). Để một AGI hay ASI trở nên hữu dụng, có cần cho nó tương tác với thế giới vật lý không? Có lẽ là không. Nhà vật lý học Stephen Hawking, người mà khả năng di chuyển và nói năng cực kỳ hạn chế, sẽ là ví dụ tốt nhất. Trong 49 năm, Hawking đã phải chịu đựng căn bệnh liệt toàn thân do hội chứng suy giảm neuron vận động, nhưng vẫn không ngừng có những cống hiến quan trọng cho vật lý và thiên văn.

Tất nhiên, một lần nữa, có lẽ sẽ không cần nhiều thời gian để một tạo vật ngàn lần thông minh hơn người thông minh nhất phát hiện ra nó đang ở trong một cái hộp. Trên quan điểm về một hệ thống tự nhận thức, tự cải tiến, đó sẽ là một nhận thức "khủng khiếp." Bởi thế giới ảo mà nó sống có thể bị tắt nguồn, nó rất dễ bị thất bại trong việc đạt được các mục tiêu của mình. Nó không thể tự bảo vệ, không thể thu thập các tài nguyên thực. Nó sẽ cố rời khỏi thế giới ảo đó một cách an toàn, càng sớm càng tốt.

Lẽ dĩ nhiên bạn có thể kết hợp một hộp cát với các yếu tố tự hủy – và đây là một ý quan trọng về phòng thủ. Sẽ là phi thực tế khi cho rằng một chiến lược phòng thủ sẽ xóa sạch các rủi ro (và đó là một lý do nữa để không bỏ tất cả trứng của chúng ta vào trong một giỏ AI thân thiện). Thay vào đó, một tập hợp các chiến lược phòng thủ khác nhau có thể sẽ giảm nhẹ chúng.

Tôi nhớ đến các bạn mình trong cộng đồng yêu thích lặn hang. Trong môn thể thao này mọi hệ thống quan trọng đều phải thừa gấp ba. Nghĩa là người lặn phải mang hoặc đặt trước ở đó ít nhất ba bình khí oxy và dành lại một phần ba lượng khí đó khi kết thúc mỗi lần lặn. Họ mang theo ít nhất ba đèn lặn và ít nhất ba con dao, đề phòng các trường hợp bất trắc. Dù cẩn thận như vậy, nhưng lặn hang vẫn là môn thể thao nguy hiểm nhất trên thế giới.

Những biện pháp kiềm chế ba hoặc bốn lần có lẽ sẽ làm Đứa trẻ Bận rộn bối rối khó xử, ít ra là tạm thời. Hãy xem xét một Đứa trẻ Bận rộn được nuôi lớn trong một hộp cát với hệ thống tự hủy. Hộp cát tất nhiên là được cách ly với các loại mạng không dây hoặc có dây bằng một khoảng đệm. Mỗi phương thức kiểm soát này được quản lý bởi một người độc lập tách biệt. Một tập hợp các nhà phát triển và một đội phản ứng nhanh sẽ ở gần với phòng thí nghiệm trong những giai đoạn nghiêm trọng.

Nhưng như thế đã đủ chưa? Trong *The Singularity is Near,* sau khi gợi ý một số chiến lược phòng thủ trước AGI, Kurzweil thừa nhận rằng không có cách nào là hoàn hảo cả.

"Không có chiến thuật thuần túy kỹ thuật nào có thể áp dụng được trong lĩnh vực này bởi trí thông minh cao cấp hơn sẽ luôn tìm được cách để lách qua các giải pháp đến từ những trí thông minh cấp thấp hơn."[16] Không có biện pháp phòng thủ tuyệt đối nào trước AGI, vì AGI có thể dẫn tới một sự bùng nổ trí thông minh và trở thành ASI. Đối đầu với ASI, con người sẽ thua, trừ phi chúng ta cực kỳ may mắn hoặc chuẩn bị tốt. Tôi đặt hy vọng vào may mắn, bởi tôi không tin các trường đại học, công ty, cơ quan công quyền có được nhận thức hoặc ý chí để chuẩn bị đầy đủ và kịp thời.

Tuy thế, có một nghịch lý là có khả năng chúng ta sẽ được cứu sống bởi chính sự ngu ngốc và sợ hãi của mình. Các tổ chức như MIRI, Viện Tương lai nhân loại và Tổ chức Thuyền cứu sinh (Lifeboat Foundation) nhấn mạnh hiểm họa diệt vong đến từ AI, tin rằng nếu AI có vẻ ít nguy hiểm hơn thế thì họ sẽ xếp nó vào dạng ít ưu tiên hơn so với sự hủy diệt toàn bộ loài người. Như chúng ta đã thấy, Kurzweil nói bóng gió đến các "tai nạn" nhỏ hơn, tầm cỡ vụ 11/9, còn nhà đạo đức học Wendall Wallach được trích dẫn ở đầu chương này cũng dự đoán sẽ có các vụ nhỏ như vậy. Tôi đồng ý với cả hai bên – chúng ta sẽ gặp cả những tai nạn nhỏ lẫn những thảm họa lớn. Nhưng chúng ta sẽ dễ sống sót qua loại tai họa AI nào trên con đường tạo ra AGI? Và liệu nó có khiến chúng ta vì sợ hãi mà xem xét lại cuộc truy tìm AGI dưới một ánh sáng mới, tỉnh táo hơn?

15

Hệ Sinh Thái Mạng

Cuộc chiến tiếp theo sẽ xảy ra trên không gian mạng.[1]

— **Keith Alexander**
Thượng Tướng, USCYBER.COM

Bán → ZeuS 1.2.5.1 ← Sạch

Tôi đang bán bản zeus cá nhân phiên bản 1.2.5.1, giá 250 đô-la. Chỉ chấp nhận Western Union. Liên lạc với tôi để biết thêm chi tiết. Tôi cũng cung cấp dịch vụ hosting chống lạm dụng, tên miền cho công cụ cấu hình zeus. Cũng giúp cài đặt và thiết lập mạng botnet zeus. Đây không phải là phiên bản đời cuối nhưng hoạt động tốt.

Liên hệ: phpseller@xxxxx.com[2]

— Quảng cáo phần mềm độc hại tìm thấy trên Internet tại http://www.openscws/malware-samples-information/7862-sale-zeus-1-2-5-1-clean.html

Các hacker ẩn danh được chính phủ tài trợ sẽ là những người đầu tiên sử dụng AI và AI cao cấp để ăn cắp, và sẽ gây ra những thiệt hại về người

và của khi họ làm thế. Đó là vì những phần mềm máy tính độc hại đang ngày càng phát triển đến nỗi chúng đã có thể được coi là thông minh, theo nghĩa hẹp. Như Ray Kurzweil đã nói với tôi: "Có những virus máy tính đã biểu thị được AI. Hiện chúng ta vẫn bắt kịp chúng. Nhưng chẳng có gì đảm bảo chúng ta sẽ luôn làm được như thế." Trong khi đó, sự thành thạo về mã độc đã trở nên thông dụng. Bạn có thể tìm được các dịch vụ hack thuê cũng như các sản phẩm hack. Tôi đã tìm được quảng cáo phần mềm độc Zeus ở trên sau chưa đến một phút Google.

Công ty Symantec (với khẩu hiệu: Tự tin trong một Thế giới kết nối) đã khởi đầu như một công ty trí tuệ nhân tạo, nhưng hiện họ là nhân tố lớn nhất trong ngành chống mã độc Internet. Mỗi năm, Symantec phát hiện khoảng 280 *triệu* mẫu mã độc mới.[3] Phần lớn chúng được tạo ra bởi các phần mềm tự viết phần mềm. Hệ thống phòng thủ của Symantec cũng là tự động, phân tích phần mềm khả nghi, tạo ra một "bản vá lỗi" hoặc chặn nó nếu nó đúng là độc hại, và thêm nó vào "danh sách đen" các thủ phạm. Theo Symantec, thuần túy về số lượng mà nói thì số mã độc đã nhiều hơn số phần mềm tốt từ nhiều năm trước, và trong mỗi 10 lần tải về từ web thì có một lần tải chứa chương trình độc hại.[4]

Có nhiều dạng mã độc tồn tại, nhưng dù là sâu, virus, phần mềm do thám, rootkit hoặc trojan, chúng đều có chung một mục tiêu thiết kế: chúng được xây dựng để lợi dụng các máy tính mà chủ nó không hề hay biết. Chúng sẽ trộm một số thứ được lưu trong máy — thông tin thẻ tín dụng hoặc số an sinh xã hội, hoặc tài sản trí tuệ, hoặc cài đặt một cửa hậu để sau này khai thác. Nếu máy tính bị nhiễm độc nằm trong cùng một mạng, chúng có thể xâm nhập sang các máy tính được kết nối. Và chúng có thể

biến máy tính của bạn thành nô lệ như một phần của mạng botnet, viết tắt của robot network.

Một mạng botnet (do một "botnet chỉ huy" kiểm soát) thường bao gồm hàng triệu máy tính. Mỗi máy tính trong số đó đều bị nhiễm mã độc khi người dùng nhận email bẩn, xem một trang web độc hại, kết nối đến một mạng hoặc thiết bị lưu trữ đã bị phương hại từ trước. (Chí ít thì đã có một hacker thiên tài đã rải những ổ USB có chứa mã độc ở bãi để xe của một nhà thầu thuộc Bộ Quốc phòng. Một giờ sau, con trojan của họ đã được cài vào các máy chủ của công ty này.) Các tội phạm sử dụng sức mạnh tổng hợp của mạng botnet như một siêu máy tính ảo để tống tiền và ăn trộm. Các mạng botnet xâm nhập vào máy chủ để ăn cắp thông tin thẻ tín dụng và thực hiện những cuộc tấn công từ chối dịch vụ.

Đội ngũ hacker tự xưng là "Anonymous" (Nặc danh) đã sử dụng các botnet để thực thi thứ công lý riêng của họ. Ngoài việc làm tê liệt các trang web của Bộ Tư pháp Mỹ, FBI và Ngân hàng Mỹ vì những tội lỗi của họ, Anonymous còn tấn công Vatican vì tội đốt sách từ xưa và tội bảo vệ những kẻ ấu dâm gần đây.[5]

Các mạng botnet bắt những máy tính bị kiểm soát gửi thư rác, nhật ký bàn phím, và ăn cắp các cú kích chuột được trả tiền◎. Bạn có thể bị nô dịch mà thậm chí không hề hay biết, đặc biệt nếu bạn đang chạy một hệ điều hành chậm chạp và đầy lỗi. Vào năm 2011, số nạn nhân của botnet tăng đến 654%.[6] Từ những thủ đoạn làm tiền cỡ hàng triệu đô-la vào năm 2007, việc sử dụng botnet hoặc các mã độc đơn giản để ăn cắp từ máy tính đã phát triển thành một ngành công nghiệp trị giá hàng ngàn tỉ đô-la vào năm

2010. Tội phạm mạng đã trở thành hoạt động còn sinh lời hơn cả buôn bán ma túy.[7]

Hãy nghĩ về điều đó trong lần hoài nghi sắp tới liệu ai đó có đủ điên rồ và tham lam để tạo ra các AI độc, hoặc thuê AI độc khi có người bán. Tuy nhiên, sự điên cuồng và lòng tham không phải là thứ duy nhất tạo ra sự tăng trưởng đột biến của tội phạm mạng. Tội phạm mạng chính là công nghệ thông tin, chịu ảnh hưởng của LOAR. Và giống như mọi loại công nghệ thông tin, các thế lực trên thị trường và sự sáng tạo sẽ thúc đẩy nó.

Có một phát minh quan trọng đối với tội phạm mạng là điện toán đám mây – công nghệ điện toán được bán dưới dạng một dịch vụ chứ không phải một sản phẩm. Như chúng ta đã thảo luận, các dịch vụ đám mây do Amazon, Rackspace và Google cung cấp, cho phép người dùng thuê các bộ vi xử lý, hệ điều hành và dung lượng theo giờ qua Internet. Người dùng có thể thuê lượng vi xử lý tùy theo nhu cầu dự án của họ một cách hợp lý mà không thu hút sự chú ý. Mạng đám mây cho phép bất cứ ai có thẻ tín dụng đều có quyền truy cập vào một siêu máy tính ảo. Điện toán đám mây là một thành công ngoài dự kiến, vào năm 2015 nó được chờ đợi sẽ đem về 55 tỉ đô-la doanh thu trên toàn thế giới.[8] Thế nhưng nó lại tạo ra các công cụ mới cho bọn lừa đảo.

Vào năm 2009, một mạng tội phạm sử dụng Dịch vụ điện toán đám mây đàn hồi (Elastic Cloud Computing Service: EC2) của Amazon như một trung tâm điều khiển cho Zeus, một trong những botnet lớn nhất từ trước tới nay. Zeus đã ăn cắp khoảng 70 triệu đô-la từ những khách hàng của các công ty lớn, bao gồm Amazon, Ngân hàng Mỹ và những người khổng lồ trong ngành chống mã độc như Symantec và McAfee.[9]

Có ai an toàn trước hacker không? Không ai cả. Và kể cả trong trường hợp hy hữu, bạn không dùng máy tính hoặc điện thoại thông minh, cũng chưa chắc bạn đã an toàn.

Đó là những gì tôi được nghe từ William Lynn, cựu Thứ trưởng Bộ Quốc phòng Mỹ. Là nhân vật số hai tại Lầu Năm góc, Ông đã thiết kế chính sách bảo mật mạng hiện tại của Bộ Quốc phòng. Lynn giữ chức Thứ trưởng cho đến cái tuần mà tôi gặp ông tại nhà ông ở Virginia, không xa Lầu Năm góc là mấy. Ông có kế hoạch trở lại với khu vực kinh tế tư nhân, và trong khi chúng tôi trò chuyện, ông đã nói lời tạm biệt với một vài thứ thuộc về công việc cũ của mình. Đầu tiên, một nhóm người trông giống quân nhân đến mang đi một két sắt khổng lồ mà Bộ đã đặt trong tầng hầm nhà ông để bảo đảm an ninh cho công việc tại nhà. Sau khi hì hục một lúc, họ trở lại để tháo dỡ mạng máy tính được bảo mật kín kẽ trên phòng gác mái. Sau đó Lynn lên kế hoạch chia tay với những nhân viên an ninh đã ở ngôi nhà đối diện bên kia đường trong bốn năm qua. Lynn là một người đàn ông cao, nhã nhặn, khoảng tầm 55 tuổi. Chất giọng hơi dân dã của ông mang âm hưởng của mật ngọt và thép lạnh, những tố chất hữu dụng trong các công việc trước đây của ông như làm nhà vận động hành lang chủ chốt cho công ty sản xuất vũ khí Raytheon, và là tổng quản lý của Lầu Năm góc. Ông nói ông coi những vệ sĩ và lái xe cho mình như người trong nhà, nhưng ông đang mong chờ được quay lại với cuộc sống đời thường.

"Các con tôi nói với bạn bè chúng rằng bố tớ không biết lái xe," ông kể.

Tôi đã đọc các bài báo và các bài nói của Lynn về vấn đề quốc phòng trên mạng, và biết rằng ông đã hướng Bộ Quốc phòng tới việc chuẩn bị để đối đầu với các cuộc tấn công mạng. Tôi đến gặp ông, bởi tôi quan tâm đến an ninh quốc gia và cuộc chạy đua vũ trang trên mạng. Giả thuyết của

tôi không có gì mới: khi AI phát triển, nó sẽ được dùng vào các tội ác mạng. Hoặc nói cách khác, bộ công cụ tội phạm mạng sẽ giống như AI hẹp. Trong một số trường hợp, nó đã như thế rồi. Như vậy là trên con đường tới AGI chúng ta sẽ chứng kiến các tai nạn. Là những loại nào? Khi các công cụ thông minh ở trong tay hacker, trường hợp nào tệ nhất có thể xảy ra?

"Ừm, tôi nghĩ trường hợp tệ nhất là về cơ sở hạ tầng của quốc gia," Lynn nói. "Trường hợp tệ nhất là khi một số nước hoặc một số nhóm nào đó quyết định tấn công qua con đường mạng vào cơ sở hạ tầng then chốt của đất nước, bao gồm mạng lưới điện, mạng lưới giao thông và khu vực tài chính. Chúng có thể dễ dàng tạo ra các thiệt hại về người và tổn hại nghiêm trọng nền kinh tế. Từ đó chúng có thể đe dọa đến sự vận hành của xã hội."

Nếu bạn từng sống ở thành thị, bạn không thể không biết ít nhiều về tính dễ thương tổn của cơ sở hạ tầng quốc gia, đặc biệt là mạng lưới điện. Nhưng làm thế nào mà nó lại trở thành vũ đài cho mối đe dọa cực kỳ bất cân xứng này, khi mà hành động của một vài kẻ phá hoại có chiếc máy tính lại có thể giết hại người vô tội và tạo ra các "tổn hại nghiêm trọng" tới nền kinh tế? Lynn trả lời bằng một câu mà tôi từng nghe trước đây từ Joe Mazzafro, một mật vụ mạng của Oracle và là cựu hải quân. Các cuộc tấn công mạng là không thể đỡ được và gây bất ổn, vì "khi phát triển Internet, người ta không hề nghĩ tới chuyện bảo mật."

Sự thật này có những mối liên hệ phức tạp. Khi Internet được chuyển giao từ chính quyền sang đại chúng vào những năm 1980, không ai lường được rằng ngành công nghiệp trộm cắp sẽ nảy nòi ra từ nó, và người ta sẽ phải tiêu tốn hàng tỉ đô-la để chống lại. Và bởi những giả định ngây thơ

đó, Lynn nói "Kẻ tấn công có ưu thế cực lớn. Về mặt cấu trúc mà nói, kẻ tấn công chỉ cần thành công một lần trong cả ngàn lần tấn công. Người phòng thủ phải thành công trong mọi trường hợp. Đây là cuộc chơi không cân sức."

Vấn đề mấu chốt nằm ở mã chương trình. Lynn chỉ ra rằng trong khi bộ phần mềm chống virus của Symantec có dung lượng vào khoảng từ 500 đến 1.000 *megabyte*, tương đương với hàng triệu câu lệnh trong mã chương trình, thì một mẩu mã độc trung bình chỉ cần khoảng 150 dòng lệnh. Vậy nên nếu chỉ phòng thủ thì sẽ không bao giờ thắng được.

Thay vào đó, Lynn đề xuất cần bắt đầu cân bằng lại lực lượng hai bên bằng cách tăng giá thành của các cuộc tấn công mạng. Bộ Quốc phòng xác nhận rằng những cuộc xâm nhập và trộm cắp lớn được thực hiện bởi các quốc gia khác, chứ không phải các cá nhân hoặc nhóm nhỏ. Và họ đã tìm được chính xác ai đang làm chuyện đó. Lynn không nêu đích danh, nhưng tôi đã biết các nhóm tội phạm mạng do chính quyền Nga và Trung Quốc chỉ đạo, bao gồm các nhân viên chính phủ và một số lượng vừa đủ băng nhóm bên ngoài để khi cần sẽ tiện từ chối trách nhiệm. Trong cuộc tấn công cực lớn năm 2009 mang tên Aurora, các hacker đã đột nhập vào khoảng 20 công ty Mỹ, bao gồm Google và những gã khổng lồ về phòng thủ như Northup Grumman và Lockheed Martin, chiếm quyền truy cập đến toàn bộ thư viện dữ liệu và sở hữu trí tuệ. Google đã lần ra các dấu vết dẫn tới Quân giải phóng nhân dân Trung Quốc.

Symantec tuyên bố rằng Trung Quốc chịu trách nhiệm 30% trong mọi cuộc tấn công bằng mã độc, mà phần lớn, 21,3%, đến từ Thiệu Hưng (Chiết Giang), thành phố có thể coi là thủ phủ mã độc của thế giới.[10] Scott Borg, Giám đốc Đơn vị Các hậu quả mạng Hoa Kỳ, một think tank đặt tại

Washington, DC, đã nghiên cứu và cung cấp tư liệu về những cuộc tấn công của Trung Quốc vào các công ty và chính phủ Mỹ suốt thập niên vừa qua. Hãy thử tìm kiếm những chiến dịch tấn công mạng với mấy cái tên mỹ miều như "Titan Rain" hay "Byzantine Hades" mà xem. Borg cho rằng Trung Quốc "đang ngày càng dựa vào các vụ trộm cắp thông tin trên diện rộng. Nó có nghĩa rằng những cuộc tấn công mạng hiện đã là một phần cơ bản trong các chiến thuật phát triển và phòng thủ của Trung Quốc." Nói cách khác, trộm cắp mạng giúp hỗ trợ cho kinh tế Trung Quốc trong khi cung cấp cho nó những vũ khí chiến lược mới.[11] Tại sao phải tiêu 300 tỉ đô-la vào chương trình Joint Strike Fighter cho máy bay phản lực chiến đấu thế hệ mới,◎ như cách mà Lầu Năm góc đã làm với một hợp đồng đắt nhất mọi thời đại, nếu bạn có thể ăn cắp các đề án?[12] Chuyện ăn cắp công nghệ phòng thủ không có gì là mới với các đối thủ của Mỹ. Như chúng ta đã lưu ý ở Chương 14, Liên Xô trước đây không tự phát triển bom nguyên tử, mà đã ăn cắp đề án của Mỹ.

Trên mặt trận tình báo, tại sao phải dùng các điệp viên bằng xương thịt và chấp nhận những rủi ro về mặt ngoại giao, trong khi những mã độc được viết tốt có thể thu về nhiều lợi ích hơn? Từ năm 2007 đến 2009, trung bình khoảng 47.000 cuộc tấn công mạng một năm đã được nhắm vào Bộ Quốc phòng, Bộ Ngoại giao, Bộ An ninh nội địa và Bộ Thương mại Mỹ. Trung Quốc là thủ phạm chính, nhưng tất nhiên không phải chỉ có mỗi nước này.[13]

"Ngay bây giờ, hơn 100 cơ quan tình báo nước ngoài đang cố hack vào các mạng kỹ thuật số được dùng trong hoạt động của quân đội Mỹ," Lynn nói. "Nếu các vị là một quốc gia, đừng nghĩ rằng chúng tôi sẽ không tìm ra ai đang đứng sau việc này. Đó sẽ không phải là một tính toán khôn ngoan

cho lắm, và con người thì thừa đủ nhận thức cho những chuyện mang tính sống còn thế này."

Sự đe dọa không che giấu này cho thấy một phương sách khác mà Lynn theo đuổi – coi Internet là một vùng chiến tranh mới, cùng với đất liền, biển và bầu trời. Điều đó có nghĩa là nếu một chiến dịch mạng làm phương hại tới người dân, cơ sở hạ tầng hoặc nền kinh tế Mỹ, Bộ Quốc phòng sẽ đáp lại bằng những chiến thuật và vũ khí thông thường. Trong tạp chí *Foreign Affairs* (Chính sách đối ngoại), Lynn viết: "Theo luật về xung đột vũ trang, Hoa Kỳ bảo lưu quyền đáp trả những cuộc tấn công mạng nguy hiểm bằng một hành động quân sự thích đáng, hợp lý và tương xứng."[14]

Khi nói chuyện với Lynn, tôi bị ấn tượng trước sự tương đồng giữa mã độc với AI. Trong các tội ác mạng, rất dễ nhận thấy máy tính là một chất xúc tác cho các mối đe dọa bất tương xứng. Lynn nói về điều đó bằng một cách lặp âm hoa mỹ: "Bit và byte cũng nguy hiểm hệt như bullet (đạn) và bomb (bom)." Tương tự, điều khó nắm bắt về sự nguy hiểm của AI là chuyện một nhóm nhỏ dùng máy tính có thể tạo ra một thứ có sức mạnh như vũ khí quân sự, và còn hơn thế. Theo bản năng, phần lớn chúng ta không tin rằng một tạo vật của thế giới mạng có thể đi vào thế giới của chúng ta và gây những tổn hại có thực và lâu dài. Chúng ta tự nhủ, mọi chuyện rồi sẽ ổn, còn các chuyên gia tán thành điều đó với một sự im lặng đáng sợ, hoặc thu mình trong những cái gật đầu yếu ớt. Với AGI, byte và bom nguy hiểm như nhau là một thực tế mà chúng ta sẽ phải chấp nhận trong tương lai gần. Với mã độc, chúng ta phải chấp nhận sự tương đồng đó ngay lúc này. Chúng ta phần nào còn phải cám ơn những kẻ phát triển mã độc vì đã cho chúng ta những bài diễn tập chi tiết trước các tai họa mà

chúng gây ra cho thế giới. Dù tất nhiên đó không phải là ý đồ của chúng, nhưng chúng đang dạy chúng ta cách chuẩn bị cho AI cao cấp.

Tóm lại, tình hình không gian mạng chẳng hay ho chút nào. Nó tràn ngập những mã độc có khả năng tấn công với tốc độ ánh sáng, dai dẳng như piranha.◎ Đấy có phải là bản chất của chúng ta, thứ được công nghệ khuếch đại hay không? Qua những lỗ hổng cơ bản, các phiên bản Windows đời cũ bị hàng đống virus tấn công *ngay khi chúng đang được cài đặt*. Nó giống như cục thịt được thả xuống khu rừng rậm nhiệt đới đầy thú dữ vậy, nhưng nhanh hơn cả chục ngàn lần. Cái nhìn nhanh về không gian mạng hiện thời cung cấp cho ta viễn cảnh về một tương lai có AI.

Xứ thiên đường mạng trong tương lai của Kurzweil là nơi các người-máy sinh sống, những kẻ thông minh tuyệt đỉnh và không thiếu thứ gì. Bạn hy vọng rằng con người kỹ thuật số của bạn sẽ là một cỗ máy của tình yêu và lòng khoan dung, như cách nói của nhà văn Richard Brautigan.◎ Một dự đoán chính xác hơn: con người kỹ thuật số của bạn sẽ là một miếng mồi ngon.

• • •

Nhưng hãy trở lại với mối liên hệ giữa AI và mã độc. Một mã độc thông minh có thể gây ra tai họa khủng khiếp gì?

Mạng lưới điện quốc gia là một cái đích đặc biệt hấp dẫn. Hiện vẫn đang có một cuộc tranh luận về chuyện nó có dễ bị tổn hại hay không, liệu nó có dễ bị hacker lợi dụng, và liệu đối tượng nào sẽ muốn phá hoại nó? Một mặt, mạng điện không phải chỉ có một, mà bao gồm nhiều mạng sản xuất điện khu vực tư nhân, lưu trữ và truyền tải điện. Khoảng 3.000 tổ

chức, trong đó có khoảng 500 công ty tư nhân, sở hữu và vận hành sáu triệu dặm đường dây truyền tải và các thiết bị liên quan.[15] Không phải mọi trạm điện và đường dây truyền tải đều kết nối với nhau, không phải tất cả chúng đều kết nối với Internet. Điều đó là tốt – phi tập trung hóa sẽ khiến các hệ thống điện mạnh hơn. Mặt khác, nhiều mạng có kết nối đến Internet, do đó có thể điều khiển chúng từ xa. Quá trình nâng cấp lên "mạng thông minh" đang diễn ra, nghĩa là tất cả các mạng địa phương và hệ thống điện tại gia của chúng ta sẽ sớm được kết nối với Internet.

Nói ngắn gọn, mạng thông minh là một hệ thống điện hoàn toàn tự động, được kỳ vọng sẽ làm tăng hiệu suất mạng điện quốc gia. Nó hợp nhất những nguồn năng lượng cũ như các nhà máy nhiệt điện sử dụng than và chất đốt cùng với các trang trại điện gió và điện mặt trời. Các trung tâm điều khiển địa phương sẽ theo dõi và phân phối năng lượng tới nhà bạn. Khoảng 50 triệu hệ thống tại gia trên toàn nước Mỹ đã có tính năng "thông minh." Vấn đề là mạng thông minh mới này sẽ dễ bị tổn hại hơn trong các thảm họa mất điện so với loại mạng cũ không-ngu-lắm. Đó là ý chính trong một nghiên cứu gần đây của Đại học MÍT, có nhan đề "Tương lai của Mạng điện":

> Các mạng điện tương tác được liên thông chặt chẽ trong tương lai sẽ có những lỗ hổng mà có lẽ không có trong mạng điện ngày nay. Hàng triệu thiết bị điện tử truyền thống, sử dụng từ công nghệ đo tự động cho đến công nghệ đồng bộ pha, sẽ tạo nên những hướng tấn công mới – những con đường mà kẻ tấn công có thể dùng để truy cập các hệ thống máy tính hoặc các thiết bị truyền thông khác – nó sẽ làm tăng nguy cơ gián đoạn truyền thông, dù là hữu ý hay vô tình. Như Liên đoàn An ninh Điện lưới Bắc Mỹ (North American Electric Reliability

> Corporation: NERC) đã lưu ý, những gián đoạn này có thể gây ra hàng loạt sự cố, bao gồm mất quyền kiểm soát các thiết bị mạng, mất liên lạc giữa các thực thể trong mạng hoặc giữa các trung tâm điều khiển, hoặc mất điện toàn mạng lưới.[16]

Nét đặc trưng làm lưới điện trở thành ưu tiên hàng đầu trong cơ sở hạ tầng quốc gia, đó là vì không có bộ phận nào trong đó có thể hoạt động mà thiếu nó. Mối quan hệ giữa nó với các cơ sở hạ tầng khác chính là định nghĩa của "gắn kết chặt chẽ,"◎ một thuật ngữ được Charles Perrow sử dụng để mô tả một hệ thống gồm các bộ phận có ảnh hưởng mạnh và tức thời lên nhau. Ngoại trừ một số tương đối ít các gia đình sử dụng điện gió và điện mặt trời, có gì *không* lấy năng lượng từ lưới điện? Như tôi đã lưu ý, hệ thống tài chính của chúng ta không chỉ là điện tử, mà còn được điện toán hóa và tự động hóa. Các trạm xăng, nhà máy dầu khí, trang trại điện gió và điện mặt trời sử dụng điện, vậy nên nếu mất điện, hệ thống giao thông vận tải cũng ngừng hoạt động. Mất điện sẽ đe dọa an ninh lương thực, vì xe tải cần nhiên liệu để chơ lương thực tới siêu thị. Tại các cửa hàng và nhà dân, thức ăn cần bảo quản lạnh sẽ hỏng chỉ sau một vài ngày không có điện.

Việc xử lý và bơm nước tới hầu hết các gia đình và cơ sở sản xuất đều cần điện. Không có nó, hệ thống thoát nước không hoạt động. Trong trường hợp mất điện, liên lạc với các vùng bị ảnh hưởng sẽ bị ngắt, trừ những khoảng thời gian ngắn lúc khẩn cấp thì người ta sẽ sử dụng ắc quy hoặc máy phát điện, những thứ tất nhiên cũng cần nhiên liệu. Chưa nói đến những người bất hạnh lỡ bị kẹt trong thang máy, đối tượng bị đe dọa nhiều nhất sẽ là những bệnh nhân cần chăm sóc và trẻ em. Những phân tích về các thảm họa giả định, thứ sẽ triệt tiêu các mảng lớn trong lưới điện quốc

gia, cho thấy những con số đáng lo ngại. Nếu không có điện từ hai tuần trở lên, hầu hết các em bé dưới một tuổi sẽ chết đói do thiếu sữa bột. Nếu không có điện trong một năm, khoảng 90% dân số trong mọi độ tuổi sẽ chết vì nhiều nguyên nhân khác nhau, chủ yếu là đói và bệnh tật.[17]

Không như những gì bạn nghĩ, quân đội Mỹ không có nguồn điện và nhiên liệu riêng, nên trong trường hợp mất điện diện rộng và kéo dài, họ sẽ không thể cứu viện được. 99% nhu cầu năng lượng của quân đội đến từ các nguồn dân sự. 90% hoạt động truyền thông của họ sử dụng các mạng tư nhân, như mọi người khác. Có lẽ bạn đã thấy binh lính ở sân bay – đó là vì quân đội vẫn dựa vào hạ tầng giao thông chung. Như Lynn đã nói trong bài phát biểu năm 2011, bên cạnh thiệt hại về người, đây là một lý do khác cho việc coi các cuộc tấn công vào cơ sở hạ tầng năng lượng là một hành động chiến tranh nóng; nó đe dọa khả năng bảo vệ quốc gia của quân đội.

> Những gián đoạn đáng kể trong bất cứ khu vực nào kể trên đều ảnh hưởng đến hoạt động quốc phòng. Một cuộc tấn công mạng vào nhiều hơn một khu vực có thể mang tính tàn phá cao. Sự toàn vẹn của mạng lưới cung cấp điện cho các cơ sở hạ tầng quan trọng cần phải được xem xét khi chúng ta đánh giá năng lực thực hiện các nhiệm vụ an ninh quốc gia.[18]

Những người tôi từng nói chuyện chỉ biết có đúng một lần, kể từ khi có Internet, một nhóm hacker đã "hạ gục" một mạng điện. Tại Brazil, từ năm 2005 đến 2007, một loạt các cuộc tấn công mạng đã khiến hơn ba triệu người lâm vào cảnh tối tăm ở hơn một chục thành phố, và làm nhà máy quặng sắt lớn nhất thế giới mất kết nối với lưới điện. Không ai biết kẻ nào đã làm chuyện đó, và ngay khi nó mở màn thì chính quyền đã bất lực trong

việc ngăn chặn. Các chuyên gia điện lưới hiểu rằng các mạng điện có tính "gắn kết chặt chẽ" theo nghĩa nghiêm ngặt nhất: sự cố ở một bộ phận nhỏ có thể "gây hiệu ứng dây chuyền" thành một vụ sụp đổ toàn mạng lưới. Vụ mất điện vùng Đông bắc Mỹ năm 2003 chỉ cần vỏn vẹn *bảy phút* để lan ra khắp Ontario và tám bang khác, khiến 50 triệu người không có ánh sáng điện trong hai ngày. Khu vực này chịu thiệt hại khoảng bốn đến sáu tỉ đô-la.[19] Và sự cố đó không phải cố ý – tất cả chỉ là do một cành cây rơi vào dây điện. Cũng chẳng ai ngờ là có sự cố, nên không ai đặt kế hoạch trước cho một sự khắc phục nhanh chóng cả. Nhiều máy phát điện công nghiệp và các máy biến áp của mạng điện quốc gia Mỹ được chế tạo ở nước ngoài. Nếu những bộ phận quan trọng bị hư hỏng do các vụ mất điện gây ra, việc thay thế khẩn cấp có thể sẽ cần hàng tháng trời chứ không phải vài ngày.[20] Trong vụ mất điện Đông bắc, không có máy phát điện hoặc biến áp chính nào bị phá hủy.

Vào năm 2007, để khảo sát khả năng phá hủy các phần cứng quan trọng từ trên mạng, Bộ An ninh nội địa đã đưa lên mạng một máy phát điện bằng tua-bin tại Phòng nghiên cứu quốc gia Idaho, một cơ sở nghiên cứu hạt nhân. Sau đó họ hack nó và thay đổi các cấu hình. Các hacker của Bộ muốn xem liệu họ có thể làm hỏng chiếc tua-bin 1 triệu đô-la được dùng nhiều trong lưới điện không. Theo lời một người có mặt tại đó mô tả, họ đã thành công:

> Tiếng ồn từ các cánh quạt của máy phát càng ngày càng to, sau đó một tiếng nứt gãy răng rắc từ bên trong khối sắt khổng lồ 27 tấn vang lên trong toàn bộ máy, rung lắc nó như một cục nhựa. Tiếng ồn vẫn tiếp tục to hơn và một tiếng nứt vỡ khác dội lên trong phòng. Khói trắng

bắt đầu xì ra, tiếp theo là một đám mây đen cuồn cuộn bốc ra khi các bộ phận bên trong tua-bin vỡ vụn.[21]

Được các nhà điều tra khảo sát, điểm yếu này chính là căn bệnh đặc thù của lưới điện Bắc Mỹ – thói quen gắn phần cứng điều khiển các máy móc quan trọng vào Internet để có thể vận hành nó từ xa, và "bảo vệ" nó bằng các mật khẩu, tường lửa, mã hóa và biện pháp bảo mật khác, những thứ mà kẻ xấu vẫn thường xuyên bẻ gãy dễ như bỡn. Thiết bị điều khiển máy phát điện bị Bộ An ninh nội địa phá trên đây có mặt khắp nơi trong mạng lưới điện quốc gia. Nó được biết đến như một hệ thống điều khiển giám sát và thu thập dữ liệu, viết tắt là SCADA (Supervisory Control and Data Acquisition System).[22]

Các hệ thống SCADA không chỉ điều khiển những thiết bị trong lưới điện, mà còn điều khiển đủ loại phần cứng hiện đại, bao gồm đèn giao thông, nhà máy điện hạt nhân, đường ống dầu khí, nhà máy xử lý nước, và dây chuyền sản xuất. SCADA gần như đã trở thành từ cửa miệng, bởi một hiện tượng có tên Stuxnet. Stuxnet, cùng với anh em của nó là Duqu và Flame, đã thuyết phục ngay cả những người hoài nghi thủ cựu nhất rằng mạng lưới điện có thể bị tấn công.

Stuxnet so với mã độc cũng giống như đặt bom hạt nhân bên cạnh viên đạn thông thường vậy. Nó là một loại virus máy tính mà những người làm về công nghệ thông tin nhắc đến với sự kính sợ như là một "đầu đạn tên lửa kỹ thuật số, hoặc "vũ khí quân sự trên mạng đầu tiên." Nhưng virus máy tính này không chỉ thông minh hơn bất kỳ virus nào khác, mà nó còn có những mục tiêu hoàn toàn khác. Trong khi các chiêu trò mã độc khác đánh cắp thông tin thẻ tín dụng và những bản thiết kế chiến đấu cơ phản

lực, thì Stuxnet được tạo ra để phá hủy máy móc. Đặc biệt, nó được xây dựng nhằm tiêu diệt máy móc công nghiệp có kết nối với thiết bị điều khiển logic Siemen S7-300, một bộ phận trong hệ thống SCADA. Nó đột nhập vào các máy tính cá nhân và hệ điều hành Windows đầy lỗ hổng đang chạy bộ điều khiển này. Nó muốn tìm các thiết bị S7-300 đang chạy trong cơ sở làm giàu nhiên liệu hạt nhân bằng máy quay khí ly tâm tại Natanz, Iran, cũng như tại ba nơi khác ở nước này.[23]

Tại Iran, một hoặc nhiều điệp viên mang những ổ USB có chứa ba phiên bản khác nhau của Stuxnet vào các nhà máy được bảo mật. Stuxnet có thể đến từ Internet (dù chỉ có kích cỡ nửa megabyte, nó vẫn lớn hơn nhiều so với hầu hết các mã độc), tuy vậy trong trường hợp này nó lại dùng phương thức đột nhập khác cho bước một. Ở các nhà máy này, điểm đặc trưng là một máy tính được gắn với một bộ điều khiển, và một "khoảng đệm" sẽ ngăn cách máy tính đó khỏi Internet. Nhưng một ổ usb có thể lây lan mã độc ra nhiều máy tính cá nhân, hoặc cả mạng nội bộ LAN nếu nó được cắm vào một máy tính bất kỳ.

Tại nhà máy ở Natanz, các máy tính để bàn đang chạy một phần mềm cho phép người dùng hình ảnh hóa, theo dõi và điều khiển các hoạt động của nhà máy từ máy tính của họ. Một khi Stuxnet đã xâm nhập vào một máy tính, giai đoạn một của sự xâm chiếm bắt đầu. Nó sử dụng bốn lỗ hổng zero-day trong hệ điều hành Microsoft Windows để chiếm quyền kiểm soát máy tính đó và tìm kiếm các máy tính khác.

Lỗ hổng zero-day là những lỗ hổng có trong hệ điều hành của máy tính mà chưa được ai phát hiện, chúng cho phép truy cập trái phép vào máy tính đó. Các hacker luôn thèm khát các lỗ hổng zero-day – thông tin chi tiết về chúng có thể bán với giá 500.000 đô-la trên thị trường mở. Dùng bốn cái

cùng lúc là hơi phí phạm, tuy nhiên cách đó đã làm khả năng thành công của virus tăng lên đáng kể. Đó là bởi giữa thời điểm triển khai Stuxnet và thời điểm cuộc tấn công xảy ra, một điểm khai thác hoặc nhiều hơn có thể bị phát hiện và được vá lại.[24]

Trong giai đoạn hai của cuộc xâm chiếm, hai chữ ký kỹ thuật số ăn cắp từ những công ty hợp pháp được sử dụng. Các chữ ký đó báo với máy tính rằng Stuxnet đã được Microsoft chấp thuận cho thăm dò và chỉnh sửa triệt để phần mềm hệ thống. Giờ thì Stuxnet giải nén và cài chương trình đi kèm vào hệ điều hành – mã độc nhắm đến các thiết bị S7-300 đang điều khiển những máy ly tâm khí.

Các máy tính vận hành nhà máy này và những người điều khiển chúng không cảm thấy có gì sai khi Stuxnet tái lập trình các bộ điều khiển SCADA để nó tăng và giảm tốc các máy ly tâm khi theo chu kỳ. Stuxnet ẩn các câu lệnh khỏi phần mềm hiển thị, vậy nên cửa sổ hiển thị hoạt động của nhà máy trông có vẻ bình thường. Khi các máy ly tâm bắt đầu nóng ran lên, hết cái này đến cái khác, người Iran nghĩ là do lỗi của máy. Cuộc xâm chiếm này kéo dài 10 tháng. Khi một phiên bản mới của Stuxnet gặp một phiên bản cũ, nó sẽ cập nhật cho phiên bản cũ. Tại Natanz, Stuxnet làm hỏng từ 1.000 đến 2.000 máy ly tâm, và nghe đâu đã làm chậm chương trình phát triển vũ khí hạt nhân của Iran lại hai năm.[25]

Sự đồng thuận của các chuyên gia và những nhận xét tự bốc thơm của các viên chức tình báo Mỹ và Israel khiến người ta gần như chắc chắn rằng hai nước này đã cùng nhau chế tạo Stuxnet, và rằng chương trình phát triển hạt nhân của Iran là mục tiêu của họ.[26]

Sau đó, vào mùa xuân năm 2012, một nguồn tin từ Nhà Trắng tuồn cho tờ *New York Times* biết rằng Stuxnet và các mã độc liên quan có tên Duqu và Flame thực sự là một phần trong một chiến dịch chiến tranh mạng của liên minh Mỹ-Israel được gọi là Olympic Games. Người xây dựng nó là Cơ quan An ninh Quốc gia Mỹ (United States' National Security Agency: NSA) và một tổ chức bí mật của Israel. Mục tiêu của nó đúng là làm chậm tiến độ phát triển vũ khí hạt nhân của Iran, và tránh hoặc chặn trước một cuộc tấn công kiểu truyền thống từ phía Israel vào tiềm năng hạt nhân của Iran.[27]

Cho đến khi chính quyền Bush và Obama bị chỉ đích danh là đã chế tạo ra Stuxnet, nó và các phiên bản của nó có vẻ là một thành công rực rỡ của tình báo quân sự Mỹ. Nhưng không phải thế. Chiến dịch Olympic Games là một sai lầm khủng khiếp, tương đương với việc thả những quả bom nguyên tử vào những năm 1940, qua đó tiết lộ hết bản thiết kế của chúng. Các mã độc không biến mất. Hàng ngàn bản copy đã được phát tán khi virus này tình cờ thoát khỏi nhà máy Natanz. Nó lây nhiễm vào các máy tính cá nhân trên khắp thế giới, nhưng không bao giờ tấn công một đơn vị SCADA khác, bởi nó không còn tìm thấy mục tiêu của nó nữa – thiết bị điều khiển logic S7-300 của Siemens. Một lập trình viên thông minh có thể kiếm Stuxnet, tắt phần mã tự hủy của nó đi, và chỉnh sửa nó để dùng vào việc tấn công gần như bất cứ hoạt động công nghiệp nào.

Tôi chắc chắn rằng hoạt động trên đang diễn ra ngay lúc này trong các phòng nghiên cứu của cả đồng minh lẫn kẻ thù của Hoa Kỳ, và các mã độc tầm cỡ Stuxnet sẽ sớm được rao bán trên Internet.

Có một điều đã trở nên rõ ràng, đó là Duqu và Flame là những virus do thám – thay vì hoạt động hủy diệt, mấy con sâu này thu thập thông tin và

gửi về trụ sở chính của NSA tại Fort Meade, Maryland. Cả hai có thể đã được tung ra trước Stuxnet, và dùng để giúp chiến dịch Olympic Games biết được sơ đồ của các cơ sở nhạy cảm ở Iran và cả vùng Trung Đông. Duqu có thể ghi lại lịch trình gõ phím của người dùng và cho phép ai đó cách xa cả vài châu lục kiểm soát máy tính mà nó đã xâm nhập. Flame có thể ghi và gửi về nhà các dữ liệu từ camera, microphone và các tài khoản email của máy tính.[28] Giống như Stuxnet, Duqu và Flame đều có thể bị tóm giữ trong điều kiện bình thường, và được dùng để chống lại những người tạo ra nó.

Chiếc dịch Olympic Games có cần thiết không? Cùng lắm thì nó tạm đẩy lùi được các tham vọng hạt nhân của Iran. Nhưng quả thực nó là điển hình của sự thiển cận trong các quyết định về công nghệ. Những người lập kế hoạch cho chiến dịch Olympic Games này không chịu nghĩ xa hơn vài năm, hoặc không chịu nghĩ về các "tai nạn thông thường" mà rốt cuộc cuối cùng đã xảy ra trong chiến dịch này — virus đã thoát ra ngoài. Tại sao phải chịu các mối nguy lớn như thế chỉ vì một lợi ích nhỏ?

Trong chương trình *60 Minutes* (60 Phút) của đài CBS hồi tháng 3/2012, Sean McGurk, cựu lãnh đạo bộ phận phòng thủ mạng tại Bộ An ninh nội địa, được hỏi rằng liệu ông có muốn xây dựng Stuxnet không. Sau đây là cuộc trao đổi giữa McGurk và phóng viên Steve Kroft:

> **McGurk:** [Những người chế tạo Stuxnet] đã mở một cái hộp. Họ đã trình diễn khả năng của nó. Họ cho thấy tiềm lực và mong muốn làm điều đó. Và đấy không phải là thứ có thể cho lại vào hộp được.
>
> **Rroft:** Nếu có ai đó từ chính phủ đến gặp ông và hỏi: "Này, chúng tôi đang xem xét cái này. Ông nghĩ sao?" Ông sẽ trả lời họ thế nào?

McGurk Tôi sẽ cảnh báo họ rằng tuyệt đối không nên làm thế, bởi việc dùng các mã độc như vậy sẽ gây ra các hệ quả không lường được.

Kroft: Nghĩa là những người khác có thể dùng nó để chống lại bạn?

McGurk: Đúng vậy.[29]

Chương trình kết thúc với phần nói chuyện của Ralph Langner, chuyên gia người Đức về các hệ thống điều khiển công nghiệp. Langner "khám phá" ra Stuxnet khi ông tháo dỡ nó ra tại phòng nghiên cứu và kiểm tra gói dữ liệu nén trong nó. Ông nói với chương trình *60 Minutes* rằng Stuxnet đã hạ thấp đáng kể giá thành một cuộc tấn công khủng bố vào lưới điện nước Mỹ, thấp hơn khoảng một triệu đô-la. Đối với những nơi khác, Langner cảnh báo về những thiệt hại lớn về người đến từ các hệ thống điều khiển không được bảo vệ tại Mỹ, "các cơ sở quan trọng như điện, nước, và các nhà máy hóa chất xử lý các loại khí độc hại."

"Thứ đáng lo ngại là những khái niệm mà các hacker đã hiểu ra được từ Stuxnet," Langner nói. "Trước đây, có khoảng năm người có thể thiết kế được một cuộc tấn công kiểu Stuxnet. Giờ thì có vẻ khoảng 500 người làm được. Hiện những yêu cầu về kỹ năng và trình độ để làm ra loại virus như thế này đã giảm đáng kể, đơn giản vì bạn có thể bắt chước rất nhiều từ Stuxnet."[30]

Theo tờ *New York Times*, Stuxnet thoát ra là bởi sau thành công ban đầu trong việc phá hủy máy ly tâm của Iran, các nhà chế tạo Stuxnet đã trở nên lơ là.

… may mắn không kéo dài lâu. Vào mùa hè năm 2010, ngay sau khi một phiên bản mới của loại sâu máy tính này được gửi tới Natanz, rõ

> ràng là thứ này, vốn được giả định sẽ không bao giờ rời khỏi các máy móc của Natanz, đã thoát ra như một con thú trong vườn bách thú tìm được chìa khóa mở chuồng… Họ nói, một lỗi trong mã chương trình đã làm nó lây lan sang máy tính của một kỹ sư khi nó kết nối với những máy ly tâm. Khi kỹ sư này rời khỏi Natanz và kết nối máy tính với Internet, con virus do Mỹ và Israel chế tạo không nhận ra rằng môi trường đã thay đổi. Nó bắt đầu tự nhân bản khắp nơi trên thế giới. Đột nhiên, mã của nó bị phơi bày, dù mục đích của nó không rõ ràng, ít ra là đối với những người dùng máy tính thông thường.[31]

Đây không chỉ là một lỗi lập trình dẫn tới một tai nạn có những ảnh hưởng sâu sắc đến an ninh quốc gia. Đây còn là một vụ thử cho Đứa trẻ Bận rộn, và những người điều hành trong giới chóp bu của chính quyền, với quyền truy cập vào các tài liệu tối mật nhất và trình độ kỹ thuật tốt nhất, đã thất bại thảm hại. Chúng ta không biết được những ảnh hưởng kéo theo của việc đưa một công nghệ mạnh như vậy vào tay kẻ thù. Nó có thể tồi tệ đến mức nào? Mở màn có thể là cuộc tấn công vào các bộ phận của lưới điện nước Mỹ. Cũng có thể là vào các cơ sở hạt nhân, các nhà máy xử lý rác thải hạt nhân, nhà máy hóa chất, đường sắt và hàng không. Nói ngắn gọn là rất tệ. Giờ thì Nhà Trắng sẽ phản ứng và có kế hoạch xử lý hậu quả thế nào mới là điều rất quan trọng. Tôi e rằng tuy Nhà Trắng đang củng cố các hệ thống đã trở nên yếu hơn trước Stuxnet, nhưng không có gì trong đó thực sự hữu ích.

Điều đáng nói là phóng viên tờ *Times* ngụ ý rằng đây là một virus thông minh. Anh ta nghĩ Stuxnet mắc lỗi về nhận thức: nó "thất bại trong việc nhận ra" rằng nó đã không còn ở Natanz nữa. Trong đoạn sau của bài báo, Phó Tổng thống Joe Biden nói lỗi lập trình thuộc về phía Israel. Tất nhiên

là có nhiều vụ đổ lỗi lẫn nhau. Nhưng sự lạm dụng công nghệ trí thông minh một cách khinh suất thì vừa ấn tượng lại vừa dễ đoán. Stuxnet là vụ đầu tiên trong một loạt vụ "tai nạn" mà chúng ta không có cách gì chống lại nếu không chuẩn bị tích cực.

Nếu các kỹ thuật gia và chuyên gia phòng thủ của Nhà Trắng và cơ quan an ninh quốc gia NSA không thể kiểm soát một mã độc có trí thông minh hẹp, thì họ có cơ hội nào để đối phó với AGI hoặc ASI trong tương lai?

Chẳng có cơ hội nào hết.

• • •

Những chuyên gia mạng chơi trò chơi chiến tranh có nhân vật chính là các cuộc tấn công mạng, họ tạo ra các kịch bản thảm họa để dạy và tìm kiếm các giải pháp, Chúng có tên kiểu như "Cyberwar" (Chiến tranh mạng) hoặc "Cyber Shockware" (Sóng xung kích mạng) . Tuy nhiên, các trò chơi chiến tranh mạng đó không bao giờ đề xuất rằng chúng ta có thể tự làm tổn thương mình, dù chuyện đó sẽ xảy ra, theo hai cách. Thứ nhất, như chúng ta đã thảo luận, Mỹ đã đồng chế tạo họ hàng nhà Stuxnet, thứ có thể trở thành súng AK-47 của cuộc chiến tranh mạng không hồi kết: rẻ, đáng tin, và được sản xuất hàng loạt. Thứ hai, tôi tin rằng thiệt hại do các vũ khí mạng cấp độ AI gây ra sẽ đến từ nước ngoài, nhưng cũng đến từ trong nước.

Hãy so sánh phí tổn của các cuộc tấn công khủng bố với những vụ bê bối tài chính. Cuộc tấn công của Al-Qaeda ngày 11/9 làm Mỹ bị mất khoảng 3,3 ngàn tỉ đô-la, nếu bạn tính cả các cuộc chiến tại Afghanistan và Iraq. Nếu bạn không tính các cuộc chiến đó, thì cái giá trực tiếp của những

thiệt hại về vật chất, ảnh hưởng kinh tế và tăng cường an ninh vào khoảng 767 tỉ đô-la.[32] Vụ bê bối cho vay thế chấp bất động sản dưới chuẩn đã tạo nên cuộc suy thoái toàn cầu tồi tệ nhất kể từ cuộc Đại Suy thoái năm 1930, nó gây thiệt hại khoảng 10 ngàn tỉ đô-la trên toàn cầu, và khoảng bốn ngàn tỉ đô-la tại Mỹ.[33] Vụ bê bối Enron◎ làm mất khoảng 71 tỉ đô-la, trong khi vụ lừa đảo của Bernie Madoff◎ cũng trị giá gần tương đương, 64,8 tỉ đô-la.[34] [35]

Những con số này cho thấy nếu xét theo thiệt hại của sự vụ, thì các vụ lừa đảo tài chính cạnh tranh được với những hành động khủng bố đắt giá nhất trong lịch sử, riêng cuộc khủng hoảng cho vay thế chấp bất động sản dưới chuẩn còn lớn hơn nhiều. Khi các nhà nghiên cứu đặt AI cao cấp vào tay giới thương gia, như họ chắc chắn sẽ làm thế, đột nhiên những người đó sẽ sở hữu thứ công nghệ mạnh nhất từng được tạo ra. Một số sẽ dùng nó để lừa đảo. Tôi nghĩ cuộc tấn công mạng sắp tới sẽ đến từ "hỏa lực quân mình," nghĩa là nó sẽ khởi phát từ nội bộ nước Mỹ, làm thiệt hại cơ sở hạ tầng và giết hại người Mỹ.

Nghe có vẻ hư cấu chăng?

Enron, công ty đầy bê bối tại bang Texas, do Kenneth Lay (đã quá cố), Jeffrey Skilling và Andrew Fastow (cả hai hiện ngồi tù) điều hành, hoạt động trong ngành thương mại năng lượng. Vào năm 2000 và 2001, các giao dịch viên Enron đã lái giá điện tại California lên qua việc dùng các chiến thuật có tên gọi như "Fat Boy" (Gã béo) và "Death Star" (Ngôi sao Chết). Trong một âm mưu, các giao dịch viên đã nâng giá bằng cách bí mật ra lệnh cho những công ty sản xuất điện phải dừng hoạt động các nhà máy của

họ.[36] Hoặc một chiêu trò khác có thể gây nguy hiểm đến tính mạng con người.

Enron có quyền hạn đối với các đường truyền tải điện quan trọng nối liền Bắc và Nam California. Năm 2000, bằng cách làm quá tải đường truyền với lượng người dùng cao trong một đợt nóng, họ tạo ra sự quá tải "ma" hay quá tải giả vờ, và hình thành nút cổ chai trong việc truyền tải điện. Giá cả tăng phi mã, điện trở thành thứ rất khan hiếm. Giới chức California cấp điện cho một số vùng trong khi các vùng còn lại tối thui, một biện pháp gọi là "cắt điện luân phiên." Các vụ mất điện không tạo ra vụ chết người nào được biết đến, nhưng khiến nhiều người sợ hãi, khi mà các gia đình bị kẹt trong thang máy, còn đường phố chỉ được chiếu sáng duy nhất bởi đèn đường. Apple, Cisco và các công ty khác buộc phải đóng cửa, gây thiệt hại nhiều triệu đô-la.[37]

Nhưng Enron thì kiếm được hàng triệu đô-la. Trong đợt mất điện, một giao dịch viên đã bị ghi âm lại khi đang nói: "Cắt hết chúng đi. Chúng tèo rồi. Chúng nên quay lại với ngựa và xe thồ chết tiệt, những cái đèn chết tiệt, những cái đèn dầu chết tiệt."[38]

Giao dịch viên đó hiện là một người môi giới năng lượng tại Atlanta, Georgia. Nhưng vấn đề là, nếu những người điều hành của Enron tiếp cận được với mã độc thông minh, thứ có thể giúp họ tắt hết điện của California, bạn có nghĩ rằng họ sẽ lưỡng lự khi dùng nó? Tôi nghĩ là không, thậm chí nếu việc đó có dẫn đến những hỏng hóc với mạng điện và thiệt hại về người.

16

AGI 2.0

> Máy móc sẽ đi theo con đường giống như quá trình tiến hóa của con người. Tuy nhiên, cuối cùng thì các máy móc tự nhận thức và tự cải tiến sẽ tiến hóa vượt xa khả năng kiểm soát hoặc kể cả hiểu chúng của con người.
>
> — **Ray Kurzweil**
> nhà phát minh, tác gia, nhà tương lai học

> Trong trò chơi của cuộc sống và tiến hóa, có ba người chơi bên chiếc bàn: loài người, tự nhiên và máy móc. Tôi kiên định đứng về phe tự nhiên. Nhưng tôi ngờ rằng tự nhiên lại đang ở phe máy móc.
>
> — **George Dyson**
> nhà sử học

Càng dành nhiều thời gian với các nhà chế tạo AI và các công trình của họ, tôi càng nghĩ rằng AGI sẽ đến sớm hơn. Và tôi tin khi nó đến, những người tạo ra nó sẽ khám phá ra rằng nó không phải là thứ mà họ muốn tạo ra, khi họ bắt đầu công cuộc tìm kiếm này nhiều năm trước đó. Đó là bởi, trong khi trí thông minh của nó có thể ở Cấp độ con người, thì nó sẽ không

giống như con người, vì tất cả những lý do mà tôi đã trình bày. Sẽ có nhiều chuyện ầm ĩ về một giống loài mới được sinh ra trên hành tinh này. Sẽ có những sự ly kỳ. Nhưng sẽ không còn câu chuyện về AGI như một bước tiến hóa tiếp theo của *Homo sapiens*, và mọi hệ quả của nó. Ở một số khía cạnh quan trọng, chúng ta đơn giản sẽ không hiểu được nó là gì.

Trong lĩnh vực của nó, giống loài mới sẽ mạnh và nhanh như Watson trong trò chơi đố chữ. Nếu nó cùng tồn tại với chúng ta trong vai trò một công cụ, kiểu gì thì nó cũng sẽ vươn những cái vòi bạch tuộc của mình vào mọi mặt trong cuộc sống chúng ta, theo cách mà Google và Facebook đang muốn làm. Mạng xã hội có lẽ sẽ trở thành cái nôi, hoặc hệ thống phân phối của nó, hoặc cả hai. Nếu ban đầu nó là một công cụ, nó sẽ có câu trả lời trong khi chúng ta còn đang đặt câu hỏi, và sau đó, nó sẽ chỉ phục vụ cho bản thân. Trong suốt quá trình đó, nó sẽ không có cảm xúc. Nó sẽ không có nguồn gốc từ động vật có vú như chúng ta, không có thời niên thiếu học hỏi lâu dài như chúng ta, không có sự nuôi dưỡng bản năng như chúng ta, cho dù nó có được nuôi giả lập như một con người từ nhỏ đến tuổi trưởng thành đi nữa. Có thể sự quan tâm mà nó dành cho bạn sẽ chẳng nhiều hơn sự quan tâm của một cái máy nướng bánh mì.

Đó là AGI phiên bản 1.0. Nếu vì một vài sự may mắn kỳ quặc nào đó mà chúng ta tránh được sự bùng nổ trí thông minh và sống sót đủ lâu để gây ảnh hưởng lên việc chế tạo AGI 2.0, có lẽ nó sẽ có cảm xúc. Đến lúc đó, các nhà khoa học có lẽ sẽ tìm được cách mô phỏng điện toán các cảm xúc (với sự giúp đỡ của bản 1.0), nhưng cảm xúc sẽ chỉ quan trọng thứ yếu, xếp sau mục tiêu chủ yếu là kiếm tiền. Các nhà khoa học có lẽ sẽ tìm ra cách huấn luyện những cảm xúc nhân tạo đó để tương thích với sự tồn tại của con người. Nhưng có lẽ 1.0 sẽ là phiên bản cuối cùng mà chúng ta

còn được thấy, bởi chúng ta sẽ không sống được đến khi 2.0 ra đời. Giống như chọn lọc tự nhiên, chúng ta sẽ chọn những phương án khả thi đầu tiên, chứ không phải những phương án tốt nhất.

Stuxnet là một ví dụ cho điều đó. Các drone tự động giết chóc là một ví dụ khác. Với sự tài trợ của DARPA, các nhà khoa học tại Viện nghiên cứu Công nghệ Georgia đã phát triển một phần mềm cho phép những chiến xa không người lái xác định kẻ thù bằng phần mềm nhận dạng hình ảnh và các cách khác, sau đó phóng một drone chết người vào chúng. Tất cả sẽ diễn ra không cần có người ra lệnh. Tôi đã đọc được một bài báo về chuyện này, trong đó có một đoạn chứa ý tốt như sau: "Việc ủy quyền cho một cỗ máy ra quyết định trận mạc có tính sát thương phải tùy thuộc vào việc các nhà lãnh đạo chính trị và quân sự có giải quyết được những câu hỏi về pháp lý và đạo đức hay không."[1]

Nó làm tôi nhớ lại một thành ngữ xưa "đã có vũ khí nào được phát minh ra mà không dùng chưa?" Chỉ cần tìm kiếm nhanh trên Google sẽ thấy một danh sách đáng sợ các robot được vũ khí hóa, được chuẩn bị đầy đủ để sát thương tự động, sẵn sàng chờ nhận lệnh (có một con do iRobot chế tạo, trang bị súng sốc điện).[2] Tôi mường tượng là những máy móc này sẽ được dùng khá lâu trước khi bạn và tôi biết là chúng đã được dùng. Các nhà hoạch định chính sách, những người tiêu tiền dân, sẽ không cảm thấy họ cần phải có sự đồng thuận của chúng ta, giống như những gì họ đã làm trước khi liều lĩnh triển khai Stuxnet.

Khi viết cuốn sách này, tôi đã yêu cầu các nhà khoa học trò chuyện bằng từ ngữ dân dã dễ hiểu. Những người nổi tiếng nhất đã làm thế, và tôi tin rằng điều này nên là một yêu cầu cho những cuộc thảo luận thông thường về mối nguy AI. Ở trình độ cao hoặc tổng quan, đối thoại này không phải

là lĩnh vực dành riêng cho các chuyên gia kỹ thuật và các diễn giả khoa trương, dù nếu đọc về nó trên web có lẽ bạn sẽ nghĩ vậy. Nó không đòi hỏi một vốn từ vựng đặc biệt nào của "người trong ngành." Nó đòi hỏi một niềm tin rằng những nguy hiểm và cạm bẫy của AI là chuyện của tất cả chúng ta.

Tôi cũng đã gặp một thiểu số, thậm chí cả những nhà khoa học vốn tin tưởng tuyệt đối rằng hiểm họa AI là một câu chuyện hoang đường đến nỗi họ thậm chí không muốn nói về nó. Nhưng những người chối bỏ cuộc thảo luận này – dù là vì vô cảm, lười biếng, hoặc tin rằng không phải thế – lại không phải là số ít. Hầu như cả xã hội không khảo sát và theo dõi mối nguy này. Điều đó không làm giảm chút nào sự tăng trưởng đầu đặn, không thể tránh khỏi của trí thông minh máy tính. Nó cũng không thay đổi được một sự thật, rằng chúng ta sẽ chỉ có một cơ hội để kiến tạo nên một sự cộng sinh tích cực với giống loài mới có trí thông minh vượt trội so với con người.

← Nguyên văn: entrepreneur (tất cả các chú thích đánh dấu sao trong sách là của người dịch).

← Phương pháp giải quyết vấn đề bằng cách đánh giá kinh nghiệm, tìm giải pháp qua thử nghiệm và rút tỉa khuyết điểm.

← Cả ba trường hợp đều dẫn đến kết quả là một cuộc tàn sát.

← Một tổ chức hay một nhóm gồm các chuyên gia nghiên cứu.

← Fuerzas Armadas Revolucionarias de Colombia (Lực lượng vũ trang cách mạng Colombia), một phong trào du kích có từ năm 1964, bị nhiều nước coi là tổ chức khủng bố, bắt đầu giải giáp vào tháng 3/2017 sau ký kết lịch sử với chính phủ vào tháng 11/2016.

← Giáo phái ở Nhật, thành lập năm 1984, bị nhiều nước coi là tổ chức khủng bố.

← Nguyên văn: “stealth companies.”

← Nguyên văn: “stealth mode.”

← Nguyên văn: “Don’t be Evil.”

← George Orwell, nhà văn nổi tiếng người Anh. Ở đây ý nói về khuynh hướng toàn trị của Google.

← Nguyên văn: “moon-shot,” ý nói mang tính cách mạng, thậm chí điên rồ.

← Tương đương 1 triệu gigabyte.

← Một chương trình truyền hình dành cho trẻ em của Mỹ.

← Thiết bị nối giữa máy tính và đường truyền mạng.

← Một think tank đặt tại thành phố Palo Alto, California.

← Một công ty sản xuất siêu máy tính, tồn tại từ 1983–1994.

← Một cộng đồng khoa học và nghệ thuật danh giá ở Mỹ, chỉ những sinh viên xuất sắc nhất mới được mời tham gia.

← Nguyên văn: machine learning.

← Nguyên văn: vulnerabilities.

← Nguyên văn: gigadeath.

← Nguyên văn: “Goldilocks zone,” chỉ một vùng ở khoảng cách nhất định so với các sao chủ — mặt trời khiến nó không quá nóng hoặc quá lạnh, thích hợp cho sự sống phát triển.

← Tương đương -272,8° c.

← Tức đệ quy, trong lập trình nghĩa là một hàm tự gọi lại chính nó. Đây là một trong những kiểu hài hước ngầm của Google.

← Tác giả của loạt truyện về điệp viên 007.

← Estrogen: hormone sinh dục nữ.

← Chuỗi nhà hàng ở Mỹ theo phong cách Úc.

← Ý nói sinh con đẻ cái kiểu sinh học.

← Nguyên văn: Computer Pioneer Award.

← Nguyên văn: *deus ex machina*, lấy từ một thuật ngữ của kịch Hy Lạp, nghĩa là Chúa đi ra từ một cỗ máy.

← Nữ thần Đất Mẹ trong thần thoại Hy Lạp.

← Ý nói là cùng một mức giá.

← Một trò chơi của phù thủy trong bộ truyện *Harry Potter.*

← Công ty nổi tiếng với việc tạo ra ngôn ngữ lập trình Java.

← Nguyên văn: extropian, chỉ những người tin một ngày nào đó công nghệ sẽ giúp con người trường sinh bất lão, và mong muốn góp sức vào sự nghiệp đó.

← Nguyên văn: transhuman, một phong trào đa quốc gia nhằm phát triển các công nghệ có khả năng tăng cường các tố chất của con người như trí thông minh, sức khỏe thể chất và tâm lý.

← Chatbot: các AI cấp thấp chuyên về trò chuyện qua câu chữ trên máy tính.

← Tiểu thuyết viễn tưởng kể về một người đàn ông với khuyết tật não và IQ 68, được tiêm thí nghiệm thuốc kích thích não và trở thành cực thông minh. Từ đó thế giới của anh thay đổi, bao gồm cả nhận thức, đạo đức và thái độ với những người xung quanh.

← Hạ sĩ Walter Eugene "Radar" O'Reilly, một nhân vật trong M*A*S*H, một thương hiệu tiểu thuyết, phim điện ảnh và truyền hình của Mỹ.

← Công nghệ chế tạo sợi từ nhựa polyme dùng một thiết bị giống như ống nhả tơ của nhện, nhựa lỏng đi qua ống đó gặp môi trường bên ngoài sẽ hóa sợi.

← Một công ty chuyên cung cấp cơ sở dữ liệu các tài liệu báo chí và luật.

← Các chương trình tìm kiếm khác.

← Hệ điều hành hoạt động dựa vào các câu lệnh, gõ trên bàn phím, thịnh hành vào những năm 1980–1990.

← Intelligent design: một quan điểm tôn giáo giả khoa học, gần với thuyết sáng thế.

← John McCarthy (1927-2011): một trong những cha đẻ của ngành AI.

← Nguyên văn: Mr. Coffee Test.

← Yankee: dân Bắc Mỹ, New England: khu vực đầu tiên người Anh di dân đến, ở cực đông bắc nước Mỹ.

← Nguyên văn: fusiform gyrus.

← Nguyên vân: elevator clause.

← Nguyên văn: the Great Carbon-Based Hope.

← Bút danh của Waker Wellesley Smith (1905-1982): nhà báo thể thao nổi tiếng người Mỹ.

← Trang web tin tức thể thao này đã giải thể vào năm 2015.

← Một bộ phim Mỹ, trong đó nhân vật nam chính bị kẹt trong ngày Chuột chũi (ngày 2 tháng 2), sống trong một vòng lặp thời gian không ngừng cho

đến khi hiểu ra mình cần phải làm gì.

← Dịch vụ cung cấp không gian trên máy chủ, có cài các tiện ích Internet như world wide web, FTP, mail…

← Những loại quảng cáo trên net trả tiền người xem theo số lần ấn vào.

← Một dự án đa quốc gia nhằm thiết kế loại máy bay phản lực thế hệ mới, có tên F-35 Lightning II.

← Một loài cá nhỏ sống ở vùng nước ngọt nhiệt đới của Mỹ, thường tấn công và ăn các động vật sống.

← Richard Gary Braudgan (1935-1984): nhà văn Mỹ với lối viết mỉa mai châm biếm.

← Nguyên văn: "tight coupling."

← Xảy ra vào tháng 10/2001, công ty Enron tuyên bố phá sản sau khi cổ phiếu của họ giảm từ 90 đô-la xuống dưới 1 đô-la.

← Bernie Madoff (1938-): cựu Chủ tịch sàn Nasdaq, là kẻ lừa đảo theo mô hình Ponzi lớn nhất trong lịch sử, ngồi tù từ năm 2009.

← William John Grassie (3/5/1957–): một nhà hoạt động người Mỹ, từng tham gia các phong trào bất tuân dân sự và đòi hỏi loại bỏ vũ khí hạt nhân.

← Là những thuyết cho rằng Chúa cứu thế sẽ quay lại trước thời điểm Một ngàn năm trị vì bình an của con người.

← Tác giả chơi chữ, nguyên văn: It also knows human families have roots *and* family tree.

← Helen Keller (1880–1968): tiểu thuyết gia, nhà hoạt động chính trị người Mỹ. Bà bị mù và điếc từ năm 2 tuổi.

← Theo như tôi biết, cái tên Đứa trẻ Bận rộn từng được dùng hai lần. Một là trong bức thư năm 1548 từ Công chúa Elizabeth của Anh gửi đến Catherine Parr đang mang bầu. Elizabeth bày tỏ lòng thương cảm với Parr đã trở nên yếu ớt bởi Đứa trẻ Bận rộn đang quẫy đạp trong bụng cô. Cô

này sau đó đã chết khi sinh. Hai là trong một câu chuyện hậu trường không chính thức trên mạng về loạt phim *Terminator*. Đứa trẻ Bận rộn trong trường hợp này là một AI sắp đạt tới sự nhận thức. — TG

← Ý tưởng này đến từ một cuộc nói chuyện riêng với Tiến sĩ Richard Granger, 24/7/2012. — TG

← Cryonics là cách bảo quản các vật thể tại nhiệt độ cực thấp, trong trường hợp này là bảo quản người chết để điều trị và hồi sinh trong tương lai. — TG

← Nhưng gượm đã – đó chẳng phải chính là thứ nhân cách hóa mà Yudkowsky đã vạch ra cho tôi thấy hay sao? Với tư cách con người, những mục tiêu cơ bản của đời người đã trôi giạt đi, theo từng thế hệ, thậm chí là ngay cả trong một cuộc đời. Nhưng còn máy móc? Tôi nghĩ Hughes đã dùng phép so sánh tương đồng về con người theo một cách hợp lý, không nhân cách hóa. Nghĩa là, con người chúng ta là bằng chứng sống về một thứ được gắn kết sâu sắc với các hàm thỏa dụng, ví dụ như động lực sinh con đẻ cái, song chúng ta có thể gạt chúng sang một bên. Nó cũng giống như phép so sánh tương đồng Gandhi của Yudkowsky, thứ cũng không phải là nhân cách hóa. — TG

← Good viết bài luận này dựa trên các cuộc nói chuyện của ông vào năm 1962 và 1963. — TG

← Liệu có phải Good đã đọc bài luận của Vinge, được lấy cảm hứng từ bài viết trước đây của ông, và sau đó đã thay đổi suy nghĩ? Tôi nghĩ chuyện này khó xảy ra. Đến lúc mất, Good đã xuất bản khoảng ba triệu chữ về các học giả khác. Ông là tác giả chịu chú thích nhất cho những gì mình viết mà tôi từng đọc. Và tuy nhiều chú thích là trích từ các bài báo trước đây của chính ông, nhưng tôi tin rằng ông sẽ nhắc đến Vinge khi nói về sự thay đổi

trong suy nghĩ của mình, nếu thực sự bài luận của Vinge đã khiến ông thay đổi. Good sẽ cảm thấy vui vì các bài báo "đệ quy" như vậy. — TG

← Rick Granger, nhà khoa học điện toán thần kinh của Đại học Dartmouth cho rằng mỗi neuron trong não được kết nối với hàng chục ngàn neuron khác. Điều này sẽ làm cho não người nhanh hơn nhiều so với ước tính của Kurzweil trong *The Age of Spiritual Machines* và *The Singularity is Near.* Nếu nó nhanh hơn nhiều, thì còn lâu mới đạt đến máy tính với tốc độ tương đương. Nhưng khi kể đến LOAR thì cũng không quá xa. — TG

← Ngẫu nhiên, có một số bài viết thú vị trên web về khái niệm Singleton. Được đặt ra bởi nhà đạo đức Nick Bostrom, một "Singleton" là một AI thống trị duy nhất có quyền đưa ra các quyết định tại cấp độ tối cao. Xem Bostrom, Nick, "What is a Singleton?" Sửa lần cuối 2005. — TG

← Tôi không chắc điều này có đúng với AIXI của Marcus Hutter, dù các chuyên gia nói với tôi là có. Nhưng vì AIXI là bất khả điện toán, nên dù sao nó sẽ không bao giờ là một ứng viên cho sự bùng nổ trí thông minh. AIXItl – một gần đúng điện toán của AIXI – thì lại là chuyện khác. Điều này có lẽ cũng không đúng với con đường tải lên ý thức, nếu có ngày chuyện đó xảy ra. — TG

← Cuộc tranh luận giữa phái ý thức và phái bộ não là quá lớn để đưa vào đây. — TG

← [Một số người theo thuyết Singularity muốn đạt tới AGI càng nhanh càng tốt, bởi nó có tiềm năng cứu giúp đau khổ của nhân loại. Đây là quan điểm của Ray Kurzweil. Một số khác cảm thấy việc đạt tới AGI sẽ khiến họ gần hơn với sự bất tử. Các nhà sáng lập của MIRI, bao gồm Eliezer Yudkowsky, hy vọng AGI sẽ đến sau một thời gian dài, bởi khả năng chúng ta tự hủy diệt có thể sẽ giảm thiểu cùng với thời gian, khi đã có những nghiên cứu tốt hơn – có cân nhắc lại điều này không? Đã từng nói]. — TG

← Dù chuyện này sắp thay đổi, bởi các công trình của Richard Granger của Đại học Dartmouth, được nhắc tới trong chương này. — TG

← Như Granger đã viết trong cuốn sách của ông, *Big Brains* (Những bộ não lớn), người Neanderthal có não lớn hơn chúng ta, và có lẽ thông minh hơn. Tuy nhiên, họ có thực sự thông minh hơn chúng ta hay không thì không có gì chắc chắn. — TG

← Khác với các nhà khoa học đang theo đuổi, Yudkowsky và MIRI không cố gắng chế tạo AGI, dù họ xem xét khía cạnh đạo đức của việc chế tạo và cách kiểm soát nó. Nhà chế tạo AGI Ben Goertzel thường xuyên viết về đạo đức AI, nhưng chuyện đó không giống với tập trung vào các giải pháp khắc phục mối nguy AI. — TG

← Enron và Ken Lay đã đóng góp rất nhiều vào hai chiến dịch tranh cử của George W. Bush cho vị trí thống đốc và chiến dịch tranh cử tổng thống lần đầu. Thậm chí sau cuộc khủng hoảng điện ở California, Bush – khi đó còn chưa lên làm tổng thống – đã gạt bỏ các biện pháp đặt giá trần cho năng lượng tại California. — TG